AOC/HXC
November 1, 2010

PATENT APPLICATION
Docket No.: 4086.1000-002

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants:	Jessica Teeling, Sigrid Ruuls, Martin Glennie, Jan G. J.van de Winkel, Paul Parren, Jørgen Petersen, Ole Baadsgaard, and Haichun Huang		
Application No.:	10/687,799	Group:	1644
Filed:	October 17, 2003	Examiner:	Ronald B. Schwadron
Confirmation No.:	1801		
For:	Human Monoclonal Antibodies Against CD20		

STATEMENT OF TOM VINK IN SUPPORT OF CORRECTION OF SEQUENCE LISTING

I, Tom Vink, Associate Director, Cell & Molecular Science at Genmab B.V. do hereby declare and state:

1. Since 1 July 2002, I am an employee of Genmab B.V., which is an affiliate of Genmab A/S.

2. I have thorough experience within the fields of molecular biology, protein biochemistry and antibody engineering.

3. The subject application is the U.S. national application equivalent to PCT application No. WO 2004/035607. I have studied the specification and the sequence listing of the PCT application prior to making this statement.

4. The subject application relates to human monoclonal antibodies against CD20, and it exemplifies three antibodies, 2F2, 7D8 and 11B8. The application contains claims relating to human monoclonal anti-CD20 antibodies, wherein the antibodies are characterized by the variable heavy chain (V_H) and variable light chain (V_L) amino acid sequences (SEQ ID Nos: 2, 6, 10, 4, 8 and 12), cf. claims 15, 17 and 19 of the application:

	2F2	7D8	11B8
V_H nucleotide	SEQ ID NO:1	SEQ ID NO:5	SEQ ID NO:9
V_L nucleotide	SEQ ID NO:3	SEQ ID NO:7	SEQ ID NO:11
V_H amino acid	SEQ ID NO:2	SEQ ID NO:6	SEQ ID NO:10
V_L amino acid	SEQ ID NO:4 (identical to SEQ ID NO:8)	SEQ ID NO:8 (identical to SEQ ID NO:4)	SEQ ID NO:12

5. I have been asked whether it would have been obvious for a person skilled in the art at the earliest priority date on 17 October 2002 (i) that the leader sequences erroneously are included in the variable heavy chain and light chain amino acid sequences (SEQ ID Nos: 2, 6, 10, 4, 8 and 12) which define the mature antibodies 2F2, 7D8 and 11B8, respectively; and (ii) in the affirmative, how the variable heavy chain and light chain amino acid sequences without leader sequences should read.

6. For the reasons explained in the following, I believe it would have been obvious for a person skilled in the art at the date of the earliest priority date on 17 October 2002 (i) that the leader sequences (marked in red in Exhibit A attached hereto) are included in the variable heavy chain and light chain amino acid sequences; and (ii) how the variable heavy chain and light chain amino acid sequences without leader sequences should read.

7. It is well known that antibodies are secretory proteins, *i.e.*, proteins that are transported to the extracellular medium by passing through the intracellular secretory pathway of the cell, the first compartment of which is the endoplasmic reticulum. Like other secretory proteins, an antibody heavy or light chain protein is initially produced as a precursor polypeptide containing a so called signal or leader peptide (also denoted leader sequence), which is necessary to direct the polypeptide into the endoplasmic reticulum for further transport and secretion. As for other soluble secretory proteins, during the transport into the endoplasmic reticulum, the signal peptides of the antibody heavy and light chains are cleaved off, resulting in a final mature protein (see also page 29, lines 35-36 of the subject application). In conclusion, any heavy or light chain from an antibody produced by *e.g.*, a hybridoma will be derived from DNA and RNA constructs in which the heavy and light chain encoding sequences are immediately preceded by the leader sequences.

8. To determine the cleavage site between the leader sequence and the antibody

sequence so as to determine where the leader sequence stops and where the antibody sequence starts, a comparison can be made with known antibody sequences. Comparison of the protein sequence (precursor polypeptide) containing the leader sequence, as deduced from the cloned RNA sequence, with a database containing known human antibody protein sequences (such as the Vbase as described on page 71, lines 7-8 of the subject application) will reveal which part of the protein is the signal peptide and where the mature protein starts. At the earliest priority date of the subject application, several of these databases were available, including the above Vbase as well as the Kabat [1] or IMGT [2] databases (see Exhibit C attached hereto for full citations of [1] and [2]).

9. More particularly, these databases contain collections of all germline signal peptides of all human antibodies, and a simple alignment of the derived amino acid sequences (V_H/V_L sequences plus leader sequences) with these signal peptide databases would reveal which part of the sequences are the signal peptides and which part of the sequences are the V_H/V_L sequences. Moreover, comparison/alignment of the derived amino acid sequences (V_H/V_L sequences plus leader sequences) with the collection of mature human V_H and V_L sequences in these databases would reveal where the mature proteins start.

As an example, in Figure 1 in Exhibit B attached hereto, a selection of V_H germline leader peptides is shown (screenshot from the Vbase). The leader peptide of V_H 7D8, MELGLSWIFLLAILKGVQC, is easily identified as sequence VH3 3-09.

Alternatively, using the DNA plot module on the Vbase site (which, to my knowledge, was available prior to the earliest priority date of 17 October 2002), it is possible to align the nucleotide sequence of your rearranged V gene to its closest mature germline V. In Figure 2 in Exhibit B this was done for the V_L region of 11B8 and, as is shown in the screenshot, the start of the mature V_L encoding sequence is gaaatt, confirming the start of the mature V_L polypeptide sequence at the corresponding amino acids EI. This can also be done with the Vquest module at the IMGT site, see Figure 3 in Exhibit B.

10. The leader sequence for a V_H sequence is typically 19-21 amino acids long, and the leader sequence for a V_L kappa sequence is typically 19-23 amino acids long. A protein always starts with a Met residue, so the leader sequences always start with a Met residue. A V_H sequence may start with a Glu or Gln residue, a V_L kappa sequence may start

with Ala, Asp, Val, Asn or Glu. V_H signal peptides can end with Cys, Ala, Ser or Pro, and V_L kappa signal peptides can end with Cys, Ala, Gly or Glu. Accordingly, a person skilled in the art would know from studying the nucleotide and amino acid sequences (SEQ ID Nos: 1-12) that the leader sequences are indeed included. Also a person skilled in the art would note that the amino acid sequences are longer than usual for the mature amino acid sequences. This information was known to a person skilled in the art at the earliest priority date of the subject application and further supports that the leader sequences are indeed included in the variable heavy chain and light chain amino acid sequences as identified by SEQ ID Nos: 2, 6, 10, 4, 8 and 12, respectively.

11. Alternatively, at the earliest priority date of the subject application several predictive methods were described, such as [3-5] (see Exhibit C attached hereto for full citations of [3-5]), by which signal peptide sequences could be predicted from a protein sequence. Using the SignalP server (which was available prior to the earliest priority date of 17 October 2002), we have performed this for the V_H polypeptide of antibody 7D8 and a screenshot of the result is shown in Figure 4 in Exhibit B attached hereto, indicating the signal peptide cleavage site between the C and E amino acid, confirming the start of the mature V_H region as EV.

12. In conclusion, it is clear that the claims defining the mature antibodies by the variable heavy chain and light chain amino acid sequences by mistake contain the leader sequences and that nothing else was intended than to define the mature antibodies without these leader sequences.

13. Monoclonal antibodies, 2F2, 7D8 and 11B8, are characterized by having the following V_H CDR1 regions in the application, cf. for example claims 33, 36 and 40 of the application:

	2F2	**7D8**	**11B8**
V_H CDR1	SEQ ID NO:13	SEQ ID NO:19	SEQ ID NO:25

The CDR regions (SEQ ID Nos: 13-18, 19-24, 25-30) are highlighted in the variable heavy and light chain sequences in Exhibit B attached hereto.

14. I have been asked whether it would have been obvious for a person skilled in the art at the earliest priority date on 17 October 2002 (i) that the V_H CDR1 regions erroneously contain an additional amino acid; and (ii) in the affirmative, how the correct V_H CDR1 regions should read.

15. I believe that it would have been obvious for a person skilled in the art at the earliest priority date on 17 October 2002 that (i) the V_H CDR1 regions contain an additional amino acid; and (ii) how the correct CDR1 regions should read for the below reasons.

16. At the earliest priority date of the subject application, several different methods were available to determine the CDR regions of antibodies. The most common method was the Kabat method [1] (see also [6] page 432-433 for an comprehensive manual for assigning the CDR regions of an antibody using the Kabat numbering scheme; see Exhibit C attached hereto for a full citation of [6]). Analysis of SEQ ID Nos. 14, 15, 16, 17, and 18 (2F2), 20, 21, 22, 23, and 24 (7D8) and 26, 27, 28, 29, 30 (11B8) defining the V_H CDR2 and CDR3 regions and the V_L CDR1, CDR2 and CDR3 regions of 2F2, 7D8 and 11B8, respectively, shows that the Kabat numbering scheme has indeed been used in this application.

17. When applying the Kabat numbering rules to the V_H CDR1, cf. [6] page 432:
"Start; Approximately residue 31 (always 9 after a C (Cys))
Residues before; always CXXXXXXXX (X meaning; any amino acid)
Residue after; Always W. Typically WV, but also WI, WA
Length; 5-7 residues "

It is obvious when applying these rules to the V_H CDR1 regions that the respective V_H CDR1 regions should be "Asp Tyr Ala Met His" for SEQ ID No 13, "Asp Tyr Ala Met His" for SEQ ID No 19 and "Tyr His Ala Met His" for SEQ ID No 25, respectively. When comparing these to the V_H CDR1 sequences as defined in the application, it appears that, by mistake, an extra amino acid has been added to these V_H CDR1 regions.

18. In conclusion, in view of the Kabat rules consistently applied in the application to designate the CDR regions, it is clear that the addition of a further amino acid to the V_H

CDR1 sequences was a mistake and that nothing else was intended than to define the V_H CDR1 sequences without inclusion of this additional amino acid.

19. I declare that all statements made in this Declaration of my own knowledge are true and that all statements made on information and belief are believed to be true. Moreover, these statements are made with the knowledge that willful false statements and the like made by me are punishable by fine or imprisonment, or both, under § 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

2 - Nov - 2010
Date

Tom Vink

Attachments:

Exhibit A: Marked-up Sequence Listing (SEQ ID NOs:1-12), color annotated
Exhibit B: Marked-up Sequence Listing (SEQ ID NOs:2, 6, and 10), color annotated; and Figures 1-4
Exhibit C: List of References cited in the above Declaration

Exhibit A

<210> 1 V_H 2F2
<211> 424
<212> DNA
<213> Homo sapiens

atggagttgg gactgagctg gattttcctt ttggctattt taaaaggtgt ccagtgtgaa 60
gtgcagctgg tggagtctgg gggaggcttg gtacagcctg gcaggtccct gagactctcc 120
tgtgcagcct ctggattcac ctttaatgat tatgccatgc actgggtccg gcaagctcca 180
gggaagggcc tggagtgggt ctcaactatt agttggaata gtggttccat aggctatgcg 240
gactctgtga agggccgatt caccatctcc agagacaacg ccaagaagtc cctgtatctg 300
caaatgaaca gtctgagagc tgaggacacg gccttgtatt actgtgcaaa agatatacag 360
tacggcaact actactacgg tatggacgtc tggggccaag ggaccacggt caccgtctcc 420
tcag 424

<210> 2 V_H 2F2
<211> 141
<212> PRT
<213> Homo sapiens

<400> 2

```
Met Glu Leu Gly Leu Ser Trp Ile Phe Leu Leu Ala Ile Leu Lys Gly
 1               5                  10                  15
Val Gln Cys Glu Val Gln Leu Val Glu Ser Gly Gly Gly Leu Val Gln
            20                  25                  30
Pro Gly Arg Ser Leu Arg Leu Ser Cys Ala Ala Ser Gly Phe Thr Phe
        35                  40                  45
Asn Asp Tyr Ala Met His Trp Val Arg Gln Ala Pro Gly Lys Gly Leu
    50                  55                  60
Glu Trp Val Ser Thr Ile Ser Trp Asn Ser Gly Ser Ile Gly Tyr Ala
65                  70                  75                  80
Asp Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Ala Lys Lys
                85                  90                  95
Ser Leu Tyr Leu Gln Met Asn Ser Leu Arg Ala Glu Asp Thr Ala Leu
            100                 105                 110
Tyr Tyr Cys Ala Lys Asp Ile Gln Tyr Gly Asn Tyr Tyr Tyr Gly Met
        115                 120                 125
Asp Val Trp Gly Gln Gly Thr Thr Val Thr Val Ser Ser
    130                 135                 140
```

<210> 3 V_L 2F2
<211> 382
<212> DNA
<213> Homo sapiens

<400> 3
atggaagccc cagctcagct tctcttcctc ctgctactct ggctcccaga taccaccgga 60
gaaattgtgt tgacacagtc tccagccacc ctgtctttgt ctccagggga aagagccacc 120
ctctcctgca gggccagtca gagtgttagc agctacttag cctggtacca acagaaacct 180

ggccaggctc ccaggctcct catctatgat gcatccaaca gggccactgg catcccagcc 240
aggttcagtg gcagtgggtc tgggacagac ttcactctca ccatcagcag cctagagcct 300
gaagattttg cagtttatta ctgtcagcag cgtagcaact ggccgatcac cttcggccaa 360
gggacacgac tggagattaa ac 382

<210> 4 V_L 2F2
<211> 127
<212> PRT
<213> Homo sapiens

<400> 4

```
Met Glu Ala Pro Ala Gln Leu Leu Phe Leu Leu Leu Leu Trp Leu Pro
 1               5                  10                  15
Asp Thr Thr Gly Glu Ile Val Leu Thr Gln Ser Pro Ala Thr Leu Ser
            20                  25                  30
Leu Ser Pro Gly Glu Arg Ala Thr Leu Ser Cys Arg Ala Ser Gln Ser
        35                  40                  45
Val Ser Ser Tyr Leu Ala Trp Tyr Gln Gln Lys Pro Gly Gln Ala Pro
    50                  55                  60
Arg Leu Leu Ile Tyr Asp Ala Ser Asn Arg Ala Thr Gly Ile Pro Ala
65                  70                  75                  80
Arg Phe Ser Gly Ser Gly Ser Gly Thr Asp Phe Thr Leu Thr Ile Ser
                85                  90                  95
Ser Leu Glu Pro Glu Asp Phe Ala Val Tyr Tyr Cys Gln Gln Arg Ser
            100                 105                 110
Asn Trp Pro Ile Thr Phe Gly Gln Gly Thr Arg Leu Glu Ile Lys
        115                 120                 125
```

<210> 5 V_H 7D8
<211> 424
<212> DNA
<213> Homo sapiens

<400> 5
atggagttgg gactgagctg gattttcctt ttggctattt taaaaggtgt ccagtgtgaa 60
gtgcagctgg tggagtctgg gggaggcttg gtacagcctg acaggtccct gagactctcc 120
tgtgcagcct ctggattcac ctttcatgat tatgccatgc actgggtccg gcaagctcca 180
gggaagggcc tggagtgggt ctcaactatt agttggaata gtggtaccat aggctatgcg 240
gactctgtga agggccgatt caccatctcc agagacaacg ccaagaactc cctgtatctg 300
caaatgaaca gtctgagagc tgaggacacg gccttgtatt actgtgcaaa agatatacag 360
tacggcaact actactacgg tatggacgtc tggggccaag ggaccacggt caccgtctcc 420
tcag 424

<210> 6 V_H 7D8
<211> 141
<212> PRT
<213> Homo sapiens

<400> 6

```
Met Glu Leu Gly Leu Ser Trp Ile Phe Leu Leu Ala Ile Leu Lys Gly
 1               5                  10                  15
Val Gln Cys Glu Val Gln Leu Val Glu Ser Gly Gly Gly Leu Val Gln
            20                  25                  30
Pro Asp Arg Ser Leu Arg Leu Ser Cys Ala Ala Ser Gly Phe Thr Phe
        35                  40                  45
His Asp Tyr Ala Met His Trp Val Arg Gln Ala Pro Gly Lys Gly Leu
    50                  55                  60
Glu Trp Val Ser Thr Ile Ser Trp Asn Ser Gly Thr Ile Gly Tyr Ala
65                  70                  75                  80
Asp Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Ala Lys Asn
                85                  90                  95
Ser Leu Tyr Leu Gln Met Asn Ser Leu Arg Ala Glu Asp Thr Ala Leu
            100                 105                 110
Tyr Tyr Cys Ala Lys Asp Ile Gln Tyr Gly Asn Tyr Tyr Tyr Gly Met
        115                 120                 125
Asp Val Trp Gly Gln Gly Thr Thr Val Thr Val Ser Ser
    130                 135                 140
```

<210> 7 V_L 7D8
<211> 382
<212> DNA
<213> Homo sapiens

<400> 7

```
atggaagccc cagctcagct tctcttcctc ctgctactct ggctcccaga taccaccgga 60
gaaattgtgt tgacacagtc tccagccacc ctgtctttgt ctccagggga aagagccacc 120
ctctcctgca gggccagtca gagtgttagc agctacttag cctggtacca acagaaacct 180
ggccaggctc ccaggctcct catctatgat gcatccaaca gggccactgg catcccagcc 240
aggttcagtg gcagtgggtc tgggacagac ttcactctca ccatcagcag cctagagcct 300
gaagattttg cagtttatta ctgtcagcag cgtagcaact ggccgatcac cttcggccaa 360
gggacacgac tggagattaa ac                                          382
```

<210> 8 V_L 7D8
<211> 127
<212> PRT
<213> Homo sapiens

<400> 8

```
Met Glu Ala Pro Ala Gln Leu Leu Phe Leu Leu Leu Leu Trp Leu Pro
 1               5                  10                  15
Asp Thr Thr Gly Glu Ile Val Leu Thr Gln Ser Pro Ala Thr Leu Ser
            20                  25                  30
Leu Ser Pro Gly Glu Arg Ala Thr Leu Ser Cys Arg Ala Ser Gln Ser
        35                  40                  45
Val Ser Ser Tyr Leu Ala Trp Tyr Gln Gln Lys Pro Gly Gln Ala Pro
    50                  55                  60
Arg Leu Leu Ile Tyr Asp Ala Ser Asn Arg Ala Thr Gly Ile Pro Ala
65                  70                  75                  80

Arg Phe Ser Gly Ser Gly Ser Gly Thr Asp Phe Thr Leu Thr Ile Ser
                85                  90                  95
Ser Leu Glu Pro Glu Asp Phe Ala Val Tyr Tyr Cys Gln Gln Arg Ser
            100                 105                 110
Asn Trp Pro Ile Thr Phe Gly Gln Gly Thr Arg Leu Glu Ile Lys
        115                 120                 125
```

<210> 9 V_H 11B8
<211> 433
<212> DNA
<213> Homo sapiens

<400> 9

```
atggagttgg ggctgagctg ggttttcctt gttgctatat taaaaggtgt ccagtgtgag 60
gttcagctgg tgcagtctgg gggaggcttg gtacatcctg gggggtccct gagactctcc 120
tgtacaggct ctggattcac cttcagttac catgctatgc attgggttcg ccaggctcca 180
ggaaaaggtc tggaatgggt atcaattatt gggactggtg gtgtcacata ctatgcagac 240
tccgtgaagg gccgattcac catctccaga gacaatgtca agaactcctt gtatcttcaa 300
atgaacagcc tgagagccga ggacatggct gtgtattact gtgcaagaga ttactatggt 360
gcggggagtt tttatgacgg cctctacggt atggacgtct ggggccaagg gaccacggtc 420
accgtctcct cag                                                   433
```

<210> 10 **V_H 11B8**
<211> 144
<212> PRT
<213> Homo sapiens

<400> 10

```
Met Glu Leu Gly Leu Ser Trp Val Phe Leu Val Ala Ile Leu Lys Gly
 1               5                  10                  15
Val Gln Cys Glu Val Gln Leu Val Gln Ser Gly Gly Gly Leu Val His
            20                  25                  30
Pro Gly Gly Ser Leu Arg Leu Ser Cys Thr Gly Ser Gly Phe Thr Phe
        35                  40                  45
Ser Tyr His Ala Met His Trp Val Arg Gln Ala Pro Gly Lys Gly Leu
    50                  55                  60
Glu Trp Val Ser Ile Ile Gly Thr Gly Gly Val Thr Tyr Tyr Ala Asp
65                  70                  75                  80
Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Val Lys Asn Ser
                85                  90                  95
Leu Tyr Leu Gln Met Asn Ser Leu Arg Ala Glu Asp Met Ala Val Tyr
            100                 105                 110
Tyr Cys Ala Arg Asp Tyr Tyr Gly Ala Gly Ser Phe Tyr Asp Gly Leu
        115                 120                 125
Tyr Gly Met Asp Val Trp Gly Gln Gly Thr Thr Val Thr Val Ser Ser
    130                 135                 140
```

<210> 11 **V_L 11B8**
<211> 382
<212> DNA
<213> Homo sapiens

<400> 11

```
atggaagccc cagcacaggt tctcttgatc ctgctactct ggctcccaga taccaccgga 60
gaaattgtgt tgacacagtc tccagccacc ctgtctttgt ctccagggga aagagccacc 120
ctctcctgca gggccagtca gagtgttagc agctacttag cctggtacca acagaaacct 180
ggccaggctc ccaggctcct catctatgat gcatccaaca gggccactgg catcccagcc 240
aggttcagtg gcagtgggtc tgggacagac ttcactctca ccatcagcag cctagagcct 300
gaagattttg cagtttatta ctgtcagcag cgtagcgact ggccgctcac tttcggcgga 360
gggaccaagg tggagatcaa ac                                          382
```

<210> 12 **V_L 11B8**
<211> 127
<212> PRT
<213> Homo sapiens

<400> 12

Met	Glu	Ala	Pro	Ala	Gln	Leu	Leu	Phe	Leu	Leu	Leu	Leu	Trp	Leu	Pro
1				5				10					15		
Asp	Thr	Thr	Gly	Glu	Ile	Val	Leu	Thr	Gln	Ser	Pro	Ala	Thr	Leu	Ser
			20					25				30			
Leu	Ser	Pro	Gly	Glu	Arg	Ala	Thr	Leu	Ser	Cys	Arg	Ala	Ser	Gln	Ser
		35					40					45			
Val	Ser	Ser	Tyr	Leu	Ala	Trp	Tyr	Gln	Gln	Lys	Pro	Gly	Gln	Ala	Pro
	50					55					60				
Arg	Leu	Leu	Ile	Tyr	Asp	Ala	Ser	Asn	Arg	Ala	Thr	Gly	Ile	Pro	Ala
65					70					75					80
Arg	Phe	Ser	Gly	Ser	Gly	Ser	Gly	Thr	Asp	Phe	Thr	Leu	Thr	Ile	Ser
				85					90					95	
Ser	Leu	Glu	Pro	Glu	Asp	Phe	Ala	Val	Tyr	Tyr	Cys	Gln	Gln	Arg	Ser
			100					105					110		
Asp	Trp	Pro	Leu	Thr	Phe	Gly	Gly	Gly	Thr	Lys	Val	Glu	Ile	Lys	
		115					120					125			

Exhibit B

<210> 2 **V_H 2F2** (Additional amino acid in V_H CDR1 region has been underlined.)
<211> 141
<212> PRT
<213> Homo sapiens

<400> 2

```
Met Glu Leu Gly Leu Ser Trp Ile Phe Leu Leu Ala Ile Leu Lys Gly
 1               5                  10                  15
Val Gln Cys Glu Val Gln Leu Val Glu Ser Gly Gly Gly Leu Val Gln
            20                  25                  30
Pro Gly Arg Ser Leu Arg Leu Ser [redacted] Ala Ala Ser Gly Phe Thr Phe
        35                  40                  45
Asn Asp Tyr Ala Met His Trp Val Arg Gln Ala Pro Gly Lys Gly Leu
    50                  55                  60
Glu Trp Val Ser Thr Ile Ser Trp Asn Ser Gly Ser Ile Gly Tyr Ala
65                  70                  75                  80
Asp Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Ala Lys Lys
                85                  90                  95
Ser Leu Tyr Leu Gln Met Asn Ser Leu Arg Ala Glu Asp Thr Ala Leu
            100                 105                 110
Tyr Tyr Cys Ala Lys Asp Ile Gln Tyr Gly Asn Tyr Tyr Tyr Gly Met
        115                 120                 125
Asp Val Trp Gly Gln Gly Thr Thr Val Thr Val Ser Ser
    130                 135                 140
```

<210> 6 **V_H 7D8** (Additional amino acid in V_H CDR1 region has been underlined.)
<211> 141
<212> PRT
<213> Homo sapiens

<400> 6

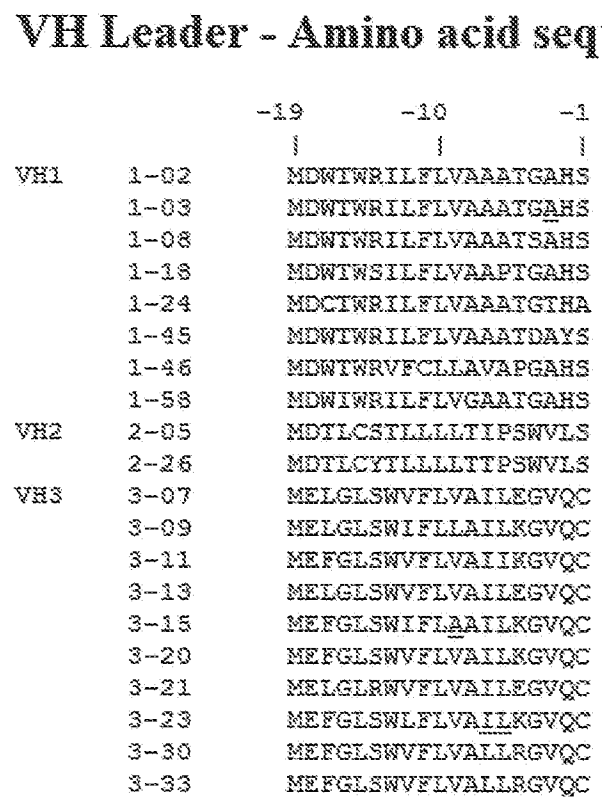
```
Met Glu Leu Gly Leu Ser Trp Ile Phe Leu Leu Ala Ile Leu Lys Gly
 1               5                  10                  15
Val Gln Cys Glu Val Gln Leu Val Glu Ser Gly Gly Gly Leu Val Gln
            20                  25                  30
Pro Asp Arg Ser Leu Arg Leu Ser [redacted] Ala Ala Ser Gly Phe Thr Phe
        35                  40                  45
His Asp Tyr Ala Met His Trp Val Arg Gln Ala Pro Gly Lys Gly Leu
    50                  55                  60
Glu Trp Val Ser Thr Ile Ser Trp Asn Ser Gly Thr Ile Gly Tyr Ala
65                  70                  75                  80
Asp Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Ala Lys Asn
                85                  90                  95
Ser Leu Tyr Leu Gln Met Asn Ser Leu Arg Ala Glu Asp Thr Ala Leu
            100                 105                 110
Tyr Tyr Cys Ala Lys Asp Ile Gln Tyr Gly Asn Tyr Tyr Tyr Gly Met
        115                 120                 125
Asp Val Trp Gly Gln Gly Thr Thr Val Thr Val Ser Ser
    130                 135                 140
```

<210> 10 **V_H 11B8** (Additional amino acid in V_H CDR1 region has been underlined.)
<211> 144
<212> PRT
<213> Homo sapiens

<400> 10

Met	Glu	Leu	Gly	Leu	Ser	Trp	Val	Phe	Leu	Val	Ala	Ile	Leu	Lys	Gly
1				5					10					15	
Val	Gln	Cys	Glu	Val	Gln	Leu	Val	Gln	Ser	Gly	Gly	Gly	Leu	Val	His
			20					25					30		
Pro	Gly	Gly	Ser	Leu	Arg	Leu	Ser	[redacted]	Thr	Gly	Ser	Gly	Phe	Thr	Phe
		35					40					45			
Ser	Tyr	His	Ala	Met	His	Trp	Val	Arg	Gln	Ala	Pro	Gly	Lys	Gly	Leu
	50					55					60				
Glu	Trp	Val	Ser	Ile	Ile	Gly	Thr	Gly	Gly	Val	Thr	Tyr	Tyr	Ala	Asp
65					70					75					80
Ser	Val	Lys	Gly	Arg	Phe	Thr	Ile	Ser	Arg	Asp	Asn	Val	Lys	Asn	Ser
				85					90					95	
Leu	Tyr	Leu	Gln	Met	Asn	Ser	Leu	Arg	Ala	Glu	Asp	Met	Ala	Val	Tyr
			100					105					110		
Tyr	Cys	Ala	Arg	Asp	Tyr	Tyr	Gly	Ala	Gly	Ser	Phe	Tyr	Asp	Gly	Leu
		115					120					125			
Tyr	Gly	Met	Asp	Val	Trp	Gly	Gln	Gly	Thr	Thr	Val	Thr	Val	Ser	Ser
	130					135					140				

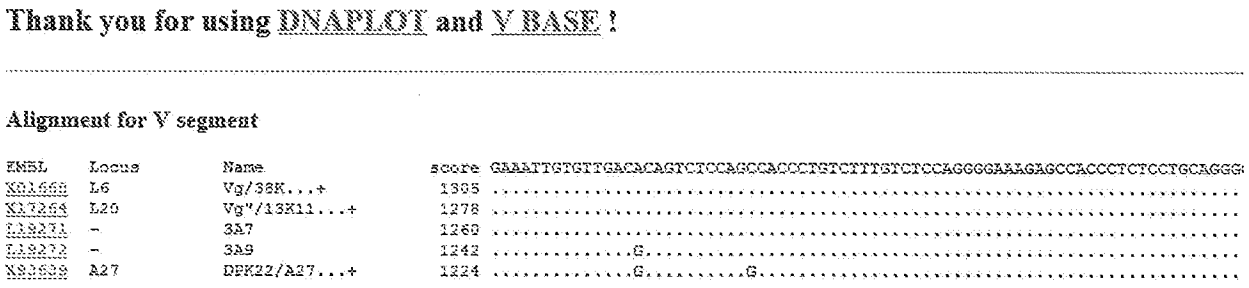
Figure 1
A screenshot is shown of part of the germline V_H signal peptides in the Vbase database.



VH Leader - Amino acid sequence alignment

		-19 -10 -1
VH1	1-02	MDWIWRILFLVAAATGAHS
	1-03	MDWIWRILFLVAAATGAHS
	1-08	MDWTWRILFLVAAATSAHS
	1-18	MDWIWSILFLVAAPTGAHS
	1-24	MDCIWRILFLVAAATGTHA
	1-45	MDWIWRILFLVAAATDAYS
	1-46	MDWTWRVFCLLAVAPGAHS
	1-58	MDWIWRILFLVGAATGAHS
VH2	2-05	MDTLCSTLLLLTIPSWVLS
	2-26	MDTLCYTLLLLTTPSWVLS
VH3	3-07	MELGLSWVFLVAILEGVQC
	3-09	MELGLSWIFLLAILKGVQC
	3-11	MEFGLSWVFLVAIIKGVQC
	3-13	MELGLSWVFLVAILEGVQC
	3-15	MEFGLSWIFLAAILKGVQC
	3-20	MEFGLSWVFLVAILKGVQC
	3-21	MELGLRWVFLVAILEGVQC
	3-23	MEFGLSWLFLVAILKGVQC
	3-30	MEFGLSWVFLVALLRGVQC
	3-33	MEFGLSWVFLVALLRGVQC

Figure 2
A screenshot of an analysis performed using the DNAplot module on the Vbase site. The complete DNA sequence of the V_L 11B8 antibody was analyzed and an alignment is shown of this sequence with the germline mature V_L regions in the database.



Thank you for using DNAPLOT and V BASE !

Alignment for V segment

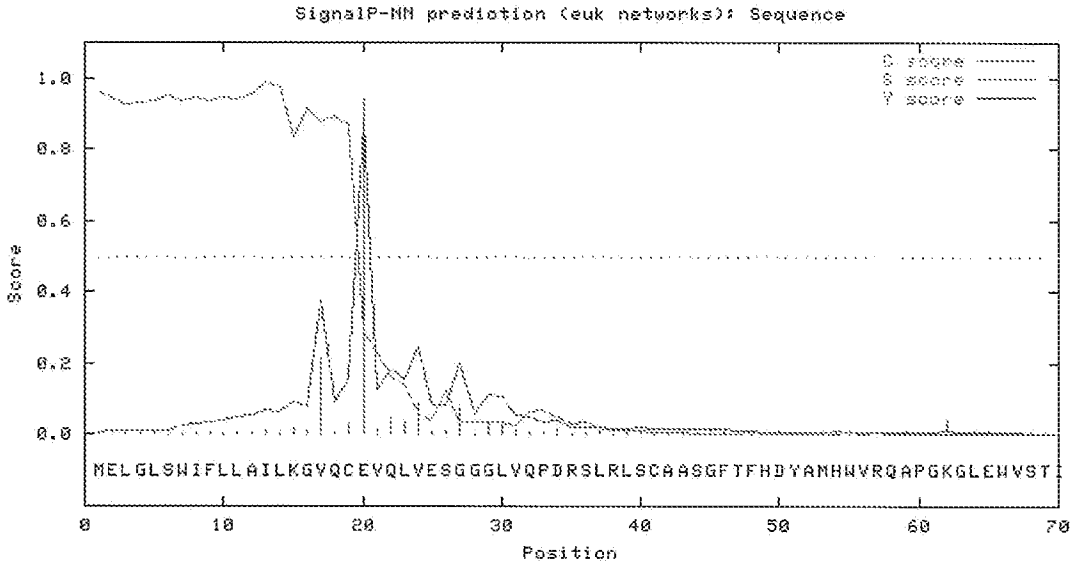
EMBL	Locus	Name	score	GAAATTGTGTTGACACAGTCTCCAGCCACCCTGTCTTTGTCTCCAGGGGAAAGAGCCACCCTCTCCTGCAGGG
X01668	L6	Vg/38K...+	1385	.....
X17264	L20	Vg"/13K11...+	1278	.....
L19271	-	3A7	1260	.....
L19272	-	3A9	1242	.....G.....
X92638	A27	DPK22/A27...+	1224	.....G.....G.....

Figure 3
A screenshot of an analysis performed using the VQUEST module on the IMGT site. The complete DNA sequence of the V_L 1188 antibody was analyzed and an alignment is shown of this sequence with the germline mature V_L regions in the database.

```
Using neural networks (NN) and hidden Markov models (HMM) trained on eukaryotes


>Sequence


SignalP-NN result:
```

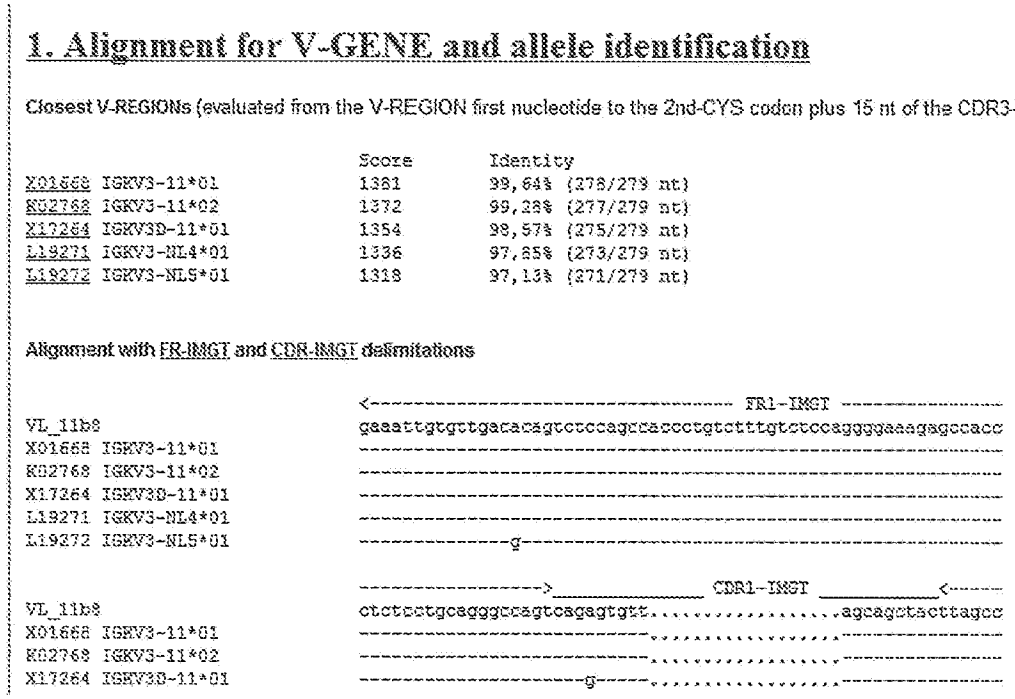



```
# data

>Sequence                 length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     20      0.940   0.32   YES
  max. Y     20      0.883   0.33   YES
  max. S     13      0.988   0.87   YES
  mean S     1-19    0.930   0.48   YES
       D     1-19    0.907   0.43   YES
# Most likely cleavage site between pos. 19 and 20: VQC-EV
```

Figure 4

The V_H polypeptide sequence of antibody 7D8 was analyzed using the SignalP server. A screenshot of the resulting prediction of the signal peptide cleavage site is shown.



1. Alignment for V-GENE and allele identification

Closest V-REGIONs (evaluated from the V-REGION first nucleotide to the 2nd-CYS codon plus 15 nt of the CDR3-IMGT)

	Score	Identity
X01668 IGKV3-11*01	1381	99,64% (278/279 nt)
K02768 IGKV3-11*02	1372	99,28% (277/279 nt)
X17264 IGKV3D-11*01	1354	98,57% (275/279 nt)
L19271 IGKV3-NL4*01	1336	97,85% (273/279 nt)
L19272 IGKV3-NL5*01	1318	97,13% (271/279 nt)

Alignment with FR-IMGT and CDR-IMGT delimitations

```
                          <---------------------------------- FR1-IMGT ------------------
VL_11b8                   gaaattgtgttgacacagtctccagccaccctgtctttgtctccaggggaaagagccacc
X01668 IGKV3-11*01        ---------------------------------------------------------------
K02768 IGKV3-11*02        ---------------------------------------------------------------
X17264 IGKV3D-11*01       ---------------------------------------------------------------
L19271 IGKV3-NL4*01       ---------------------------------------------------------------
L19272 IGKV3-NL5*01       ----------g----------------------------------------------------

                          ------------------>_____________ CDR1-IMGT _____________<------
VL_11b8                   ctctcctgcagggccagtcagagtgtt.....................agcagctacttagcc
X01668 IGKV3-11*01        ----------------------------.....................----------------
K02768 IGKV3-11*02        ----------------------------.....................----------------
X17264 IGKV3D-11*01       ----------------------g-----.....................----------------
```

Exhibit C

References cited:

1. Kabat, E.A., et al., *Sequences of Proteins of Immunological Interest, NIH Publication No. 91-3242.* Bethesda, MD: US Department of Health and Human Services, 1991.
2. Lefranc, M.P., *IMGT, the international ImMunoGeneTics database.* Nucleic Acids Res, 2001. **29**(1): p. 207-9.
3. Chou, K.C., *Prediction of signal peptides using scaled window.* Peptides, 2001. **22**(12): p. 1973-9.
4. Chou, K.C., *Using subsite coupling to predict signal peptides.* Protein Eng, 2001. **14**(2): p. 75-9.
5. Nielsen, H., et al., *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.* Protein Eng, 1997. **10**(1): p. 1-6.
6. Kontermann, R. and S. Dübel, *Antibody engineering*. 2001: Springer Heidelberg.

Kabat citation, copy available at the following website:

http://books.google.nl/books?id=3jMvZYW2ZtwC&printsec=frontcover&dq=Sequences+of+Proteins+of+Immunological+Interest&source=bl&ots=PcPHQOzm0v&sig=KWw6qWtj3dGjZ-QzDIqIanLr3Ss&hl=nl&ei=qqHQTKK9FcidOo_atOwE&sa=X&oi=book_result&ct=result&resnum=1&ved=0CBsQ6AEwAA#v=onepage&q&f=false

Alternatively, the Kabat Rules are set forth in Exhibit C (6).

Nucleic Acids Research, 2001, Vol. 29, No. 1 207–209

IMGT, the international ImMunoGeneTics database

Marie-Paule Lefranc*

Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received September 27, 2000; Revised and Accepted October 17, 2000

ABSTRACT

IMGT, the international ImMunoGeneTics database, freely available at http://imgt.cines.fr:8104, was created in 1989 at the Université Montpellier II, CNRS, Montpellier, France, and is a high quality integrated information system specialising in immunoglobulins, T cell receptors and major histocompatibility complex molecules of human and other vertebrates. IMGT provides researchers and clinicians with a common access to all nucleotide, protein, genetic and structural immunogenetics data. This information is of high value for medical and veterinary research, biotechnology related to antibody and T cell receptor engineering, genome diversity and evolution studies of the immune response.

INTRODUCTION

IMGT, the international ImMunoGeneTics database (http://imgt.cines.fr:8104) (1,2), created in 1989 at the Université Montpellier II, CNRS, Montpellier, France, is a high quality integrated information system specialising in Immunoglobulins (Ig), T cell Receptors (TcR) and Major Histocompatibility Complex (MHC) molecules of human and other vertebrates. IMGT provides common access to expertly annotated data on the genome, proteome, genetics and structure of the Ig, TcR and MHC. Due to its high quality and easy data distribution, IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas, myelomas), therapeutic approaches, biotechnology related to antibody engineering, veterinary research, genome diversity and genome evolution studies of the immune response. IMGT consists of databases ('IMGT sequence databases'), Web resources ('IMGT Marie-Paule page') and interactive tools (Fig. 1).

IMGT SEQUENCE DATABASES

The IMGT sequence databases comprise at present (i) IMGT/LIGM-DB, a comprehensive database of 41 248 Ig and TcR nucleotide sequences from human and from 104 other vertebrate species, with translation for fully annotated sequences, created by LIGM (Laboratoire d'ImmunoGénétique Moléculaire, Montpellier, France) (1–3) and (ii) IMGT/HLA-DB, a database of the 1257 human MHC allele sequences, developed by ICRF (Imperial Cancer Research Fund, Oxford, UK) and ANRI (Anthony Nolan Research Institute, London, UK) (4).

IMGT MARIE-PAULE PAGE

The IMGT Marie-Paule page comprises Web resources recorded in HTML pages (2780 documents, 3698 internal and external hyperlinks).

The IMGT Scientific chart provides the controlled vocabulary and the annotation rules and concepts defined by IMGT for the identification, description and classification of the Ig and TcR data of all vertebrate species (5). The concepts of classification have been used to set up a unique nomenclature of Ig and TcR genes (6,7), which has been adopted by the HUGO nomenclature committee. The complete list of the IMGT human Ig and TcR gene names has been entered in GDB and LocusLink (1999). A uniform numbering system for Ig and TcR sequences of all species has been established to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (Ig or TcR), chain type or species (8). IMGT has developed a formal specification of the terms to be used in the domain of immunogenetics and bioinformatics to ensure accuracy, consistency and coherence in IMGT. This has been the basis of the IMGT-ONTOLOGY (5), the first ontology in the domain, which allows the management of the immunogenetics knowledge for all vertebrate species. Control of coherence in IMGT combines data integrity control and biological data evaluation (9,10).

The IMGT Repertoire, the global Web resource in ImMunoGeneTics for the immunoglobulins and T cell receptors of human and other vertebrates, based on the 'IMGT Scientific chart', provides an easy-to-use interface to carefully expertised data on the genome, proteome, polymorphism and structure of the Ig and TcR (2). Genome data include chromosomal localisations, locus representations and germline gene tables. Proteome and polymorphism data are represented by protein displays, alignments of alleles and tables of alleles. These data are regularly published in the IMGT Locus in Focus section of *Experimental and Clinical Immunogenetics* (IMGT Index>IMGT Locus in Focus at http://imgt.cines.fr:8104). Structural data comprise 2D graphical representations designated as Colliers de Perles (1) and 3D representations of Ig and TcR variable regions (2,3). This visualisation permits rapid correlation between protein sequences and 3D data retrieved from the Protein Data Bank (PDB). A new section, currently being developed, contains data on probes used for the analysis of Ig and TcR gene rearrangements and expressions, and RFLP (Restriction Fragment Length Polymorphism) studies.

*Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr

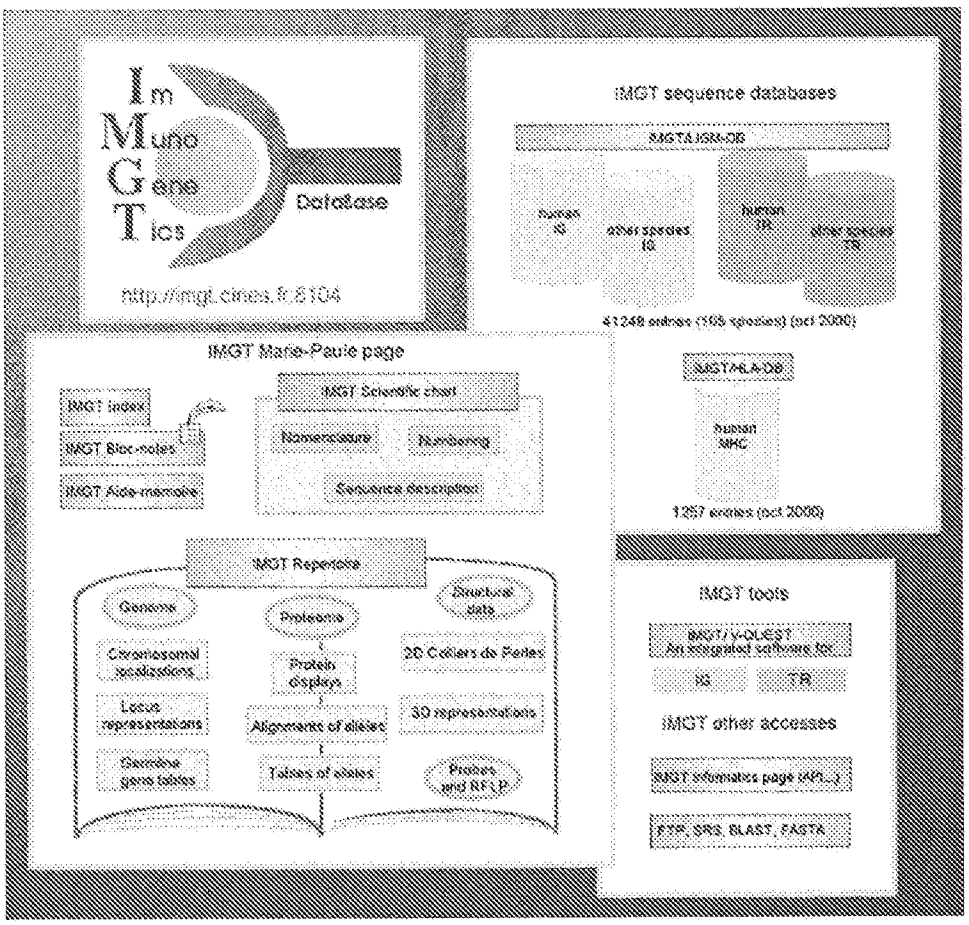


Figure 1. Overview of the IMGT information system.

The IMGT Index and the IMGT Aide-mémoire represent useful biological resources for students and researchers. The IMGT Bloc-notes provides numerous hyperlinks to the Web servers specialising in immunology, genetics, molecular biology and bioinformatics (11).

IMGT TOOLS AND OTHER ACCESSES

IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software for Ig and TcR. This tool analyses the input Ig or TcR variable nucleotide sequence and provides the nucleotide alignment by comparison with the IMGT reference directory, the protein translation of the input sequence, and its 2D Colliers de Perles representation. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers.

Since July 1995, IMGT has been available on the Web at http://imgt.cines.fr:8104. IMGT provides biologists with an easy to use and friendly interface. From January 1996 to October 2000, the IMGT WWW server at Montpellier was accessed by more than 95 000 sites. IMGT has an exceptional response with more than 6000 requests a week. IMGT data are also distributed by EBI (distribution of CD-ROM, network fileserver: netserv@ebi.ac.uk, and anonymous FTP server), and from many SRS (Sequence Retrieval System) sites. To facilitate the integration of IMGT data into applications developed by other laboratories, we have built an API (Application Programming Interface) to access the database and its software tools (10).

ELECTRONIC AND MAILING ADDRESSES

IMGT home page: http://imgt.cines.fr:8104 (IMGT contact lefranc@ligm.igh.cnrs.fr).
IMGT page at EBI: http://www.ebi.ac.uk/imgt, ftp://ftp.ebi.ac.uk/pub/databases/imgt
IMGT/LIGM-DB: http://imgt.cines.fr:8104 (contacts lefranc@ligm.igh.cnrs.fr, giudi@ligm.igh.cnrs.fr), ftp://imgt.cines.fr/IMGT (contact denys.chaume@igh.cnrs.fr).
IMGT/HLA-DB: http://www.ebi.ac.uk/imgt/hla/ (contacts jrobinso@ebi.ac.uk, julia@icrf.icnet.uk, marsh@icrf.icnet.uk).

CITING IMGT

Authors who make use of the information provided by IMGT should cite this article as a general reference for the access to and content of IMGT, and quote the IMGT home page URL, http://imgt.cines.fr:8104.

ACKNOWLEDGEMENTS

I am deeply grateful to the IMGT team for its expertise and motivation. IMGT is funded by the European Union's 5th PCRDT programme QLG2-2000-01287, the CNRS (Centre National de la Recherche Scientifique), the Ministère de l'Education Nationale and the Ministère de la Recherche. Subventions have been received from ARC (Association pour la Recherche sur le Cancer) and the Région Languedoc-Roussillon.

REFERENCES

1. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbié,V. and Chaume,D. (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **27**, 209–212.
2. Ruiz,M., Giudicelli,V., Ginestoux,C., Stoehr,P., Robinson,J., Bodmer,J., Marsh,S.G., Bontrop,R., Lemaitre,M., Lefranc,G., Chaume,D. and Lefranc,M.-P. (2000) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **28**, 219–221.
3. Lefranc,M.-P. (2000) IMGT ImMunoGeneTics Database. *International BIOforum*, **4**, 98–100.
4. Robinson,J., Malik,A., Parham,P., Bodmer,J.G. and Marsh,S.G.E. (2000) IMGT/HLA Database – a sequence database for the human major histocompatibility complex. *Tissue Antigens*, **55**, 280–287.
5. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, **12**, 1047–1054.
6. Lefranc,M.-P. (2000) Locus maps and genomic repertoire of the human Ig genes. *Immunologist*, **8**, 80–87.
7. Lefranc,M.-P. (2000) Locus maps and genomic repertoire of the human T-cell receptor genes. *Immunologist*, **8**, 72–79.
8. Lefranc,M.-P. (1999) The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *Immunologist*, **7**, 132–136.
9. Giudicelli,V., Chaume,D., Mennessier,G., Althaus,H.H., Müller,W., Bodmer,J., Malik,A. and Lefranc,M.-P. (1998) IMGT, the international ImMunoGeneTics database: a new Design for Immunogenetics Data Access. In Cesnik,B. *et al.* (eds), *Proceedings of the Ninth World Congress on Medical Informatics, MEDINFO' 98*. IOS Press, Amsterdam, 351–355.
10. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (1998) IMGT/LIGM-DB: A systematized approach for ImMunoGeneTics database coherence and data distribution improvement. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISBM-98*, 59–68.
11. Lefranc,M.-P. (2000) Web sites of Interest to Immunologists. *Curr. Protocols Immunol.*, A.1J.1–A.1J.33.

ELSEVIER

Peptides 22 (2001) 1973–1979

PEPTIDES

Prediction of signal peptides using scaled window

Kuo-Chen Chou*

Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo, MI 49007-4940, USA

Received 29 May 2001; accepted 29 June 2001

Abstract

Cells use a ZIP code system to sort newly synthesized proteins and deliver them wherever they are needed: into different internal compartments called organelles or even out of the cell altogether. One of the most essential features of the ZIP code system is the signal sequence or "address tag," which is originally present in the N-terminal part of the protein and is trimmed away by the time it is secreted. Owing to the importance of signal peptides for understanding the molecular mechanisms of genetic diseases, reprogramming cells for gene therapy, and constructing new drugs for correcting a specific defect, it is highly desirable to develop a fast and accurate method to identify the signal peptides. In this paper, a scaled window model is proposed. Based on such a model as well as Markov chain theory, a new algorithm is formulated for predicting the signal peptides. Test results for the 1939 secretory proteins and 1440 non-secretary proteins have indicated that the new algorithm is particularly successful in the overall success rate, and hence can serve as a complementary tool to the existing algorithms for signal peptide prediction.

Keywords: "Zipcode" sequence; Markov chain; Secretory proteins; Non-secretory proteins; Discriminant function

1. Introduction

Proteins with various functions are constantly being made within cells. These nascent proteins have to be transported either out of the cell, or to the different organelles within the cell. How are they transported across the membrane surrounding the organelles? And how are they directed to their correct location? These questions have been answered through the work of Günter Blobel, the Nobel Laureate of last year in Physiology or Medicine [13]. He and his co-workers have discovered that newly synthesized proteins have an intrinsic signal peptide, functioning as "address tags" or "zip codes," that is essential for direct them wherever they are needed.

The discovery of signal peptides has had an immense impact on modern cell biological research. Knowledge of signal peptides has helped explain the molecular mechanisms behind several genetic diseases. When a cell divides, large amounts of proteins are being made and new organelles are formed. If a sorting signal in a protein is changed, the protein could end up in a wrong cellular location and cause varieties of diseases. For example, in some forms of familial hypercholesterolemia, a very high level of cholesterol in the blood is due to deficient transport signals. Also, hereditary diseases, such as cystic fibrosis, are caused by the fact that proteins do not reach their proper destination. Knowledge of signal peptides will increase our understanding of processes leading to disease and hence can be used to develop new therapeutic strategies.

Today some drugs have already been produced in the form of proteins, e.g. growth hormone, insulin, and hemoglobin. Usually bacteria are used for the production of protein drugs. However, in order to be functionally proper, it is necessary to synthesize human proteins in more complex cells, such as yeast cells. The contemporary gene technology allows us to generate the genes of the desired proteins with sequences coding for transport signals. The cells with the modified genes can be efficiently used as "protein factories." Accordingly, knowledge of protein signals can then be used to reprogram cells in a specific way for future cell and gene therapy. Actually, protein signals have become a crucial tool for researchers to construct new drugs that are targeted to a particular organelle to correct a specific defect. For example, by adding a specific tag to the desired proteins, one can, for instance, tag them for excretion, making them much easier to harvest [13].

Actually, the identification of signal peptides has become a prerequisite in order to use such a technology effectively.

*Tel.: +1-616-833-8867; fax: +1-616-372-8766.
E-mail address: kuo-chen.chou@am.pnu.com (K.-C. Chou).

0196-9781/01/$ – see front matter
PII: S0196-9781(01)00540-X

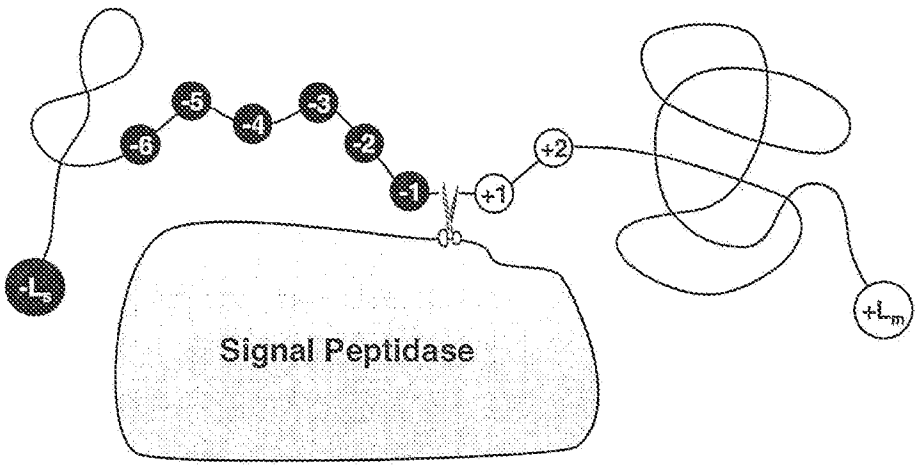


Fig. 1. A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a black circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with a black number. The signal sequence contains L_S residues, and the mature protein L_m residues. The cleavage site is at the position (−1, +1), i.e. between the last residue of the signal sequence and the first residue of the mature protein.

However, since the number of protein sequences entering into data banks has been rapidly increasing, it is time-consuming and costly to identify the signal peptides entirely by experiments. For example, the yearly increment of sequence entries in SWISS-PROT [1] in 1987 was 1,266, and that in 1988 was 3,497, but that in 1997 was already 10,092. In view of this, it is highly desirable to develop an automated algorithm to identify signal peptides of newly synthesized proteins. The existing methods for predicting the signal peptides are based mostly on the use of neural networks (see, e.g. [11,16]). As pointed out by King [14], the advantages of neural network prediction methods are: (1) "readily available," and (2) "often successful in practice"; the disadvantages are: (1) "very poor explanatory power," (2) "little use of chemical or physical theory," and (3) "statistically rather poorly characterized." Besides, although the computational costs for training the networks was considerably higher, the prediction accuracy thus obtained was not always higher (and sometimes even lower) than the analytical methods. The current study was initiated in an attempt to develop an automated analytical method to predict signal peptides based on a scaled window model and Markov chain theory (see, e.g. Bhat, [2]).

2. Materials and methods

Signal peptides comprise the N-terminal part of the secretory protein chain. They control the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes [12,18,21], and are cleaved off by signal peptidase (Fig. 1) while the protein is translocated through the membrane. As shown in Fig. 1, the cleavage site is at the sequential position between residues −1 and +1. Accordingly, the prediction of signal peptide is directly correlated with the prediction of the secretion-cleaved site. The length of signal peptides is varied for different secretory proteins. As shown in Fig. 2, of the 1939 signal peptides studied by Nielsen et al. [17], one (the shortest) contains 8 amino acid residues, one (the longest) contains 90 residues, and the majority is within the range of 18–25 residues. The extreme variation in length and sequence has made it a herculean task to formulate a general algorithm for predicting the signal peptides. To deal with this kind of situation, let us consider a window with a scale of $-\xi_1, \ldots, -3, -2, -1, +1, +2, \ldots, \xi_2$ (Fig. 3). Such a window is called "scaled window" and symbolized as $[-\xi_1, +\xi_2]$. When sliding the scaled window $[-\xi_1, +\xi_2]$ along a sequence of n residues, one can see $n - (\xi_1 + \xi_2) + 1$ different sequences. Of the sequence segments thus generated, only the one with the residue at the scale -1 being the very last residue of the signal sequence is deemed as the secretion-cleavable segment (Fig. 3a), while all the other segments deemed as non-secretion-cleavable (Fig. 3b and c). By this way, if sliding the scaled window $[-\xi_1, +\xi_2]$ along a protein sequence of n residues, one can generate one, and only one, secretion-cleavable segment and $n - (\xi_1 + \xi_2)$ non-secretion-cleavable segments if the protein is secretory; but $n - (\xi_1 + \xi_2) + 1$ non-secretion-cleavable segments if it is non-sectertory. All the secretion-cleavable segments form a positive set denoted by S^+, and all the non-secretion-cleavable segments form a negative set S^-

Segments generated by sliding the scaled window $[-\xi_1, +\xi_2]$ along protein sequences can be generally expressed as

$$R_{-\xi_1}R_{-(\xi_1-1)}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+(\xi_2-1)}R_{+\xi_2} \quad (1)$$

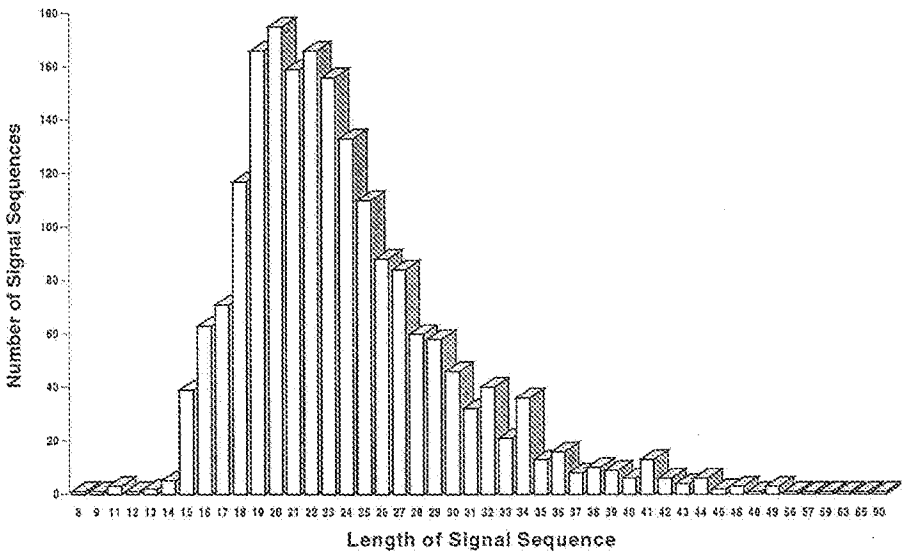


Fig. 2. A histogram to show the distribution of signal sequences with their length in the 1939 secretory proteins studied by Nielsen et al. [17].

where $R_{-\xi_1}$ represents the residue at the scale $-\xi_1$, R_{-1} the residue at the scale -1, R_{+1} the residue at the scale $+1$, and so forth.

If the amino acid residue at each of the segment subsites (eq.1) can be treated as an independent element, i.e. there is no coupling at all among these subsites, then its attribute to the cleavable set S^+ and that to the non-cleavable set S^- can be formulated, respectively, as [4]

$$\Psi_0^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^+_{-\xi_1}(R_{-\xi_1})\cdots P^+_{-3}(R_{-3})P^+_{-2}(R_{-1})P^+_{+1}(R_{+1}) P^+_{+2}(R_{+2})\cdots P^+_{+\xi_2}(R_{+\xi_2}) \quad (2a)$$

and

$$\Psi_0^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^-_{-\xi_1}(R_{-\xi_1})\cdots P^-_{-3}(R_{-3})P^-_{-2}(R_{-1})P^-_{+1}(R_{+1}) P^+_{+2}(R_{+2})\cdots P^+_{+\xi_2}(R_{+\xi_2}) \quad (2b)$$

where $P_i^+(R_i)$ is the probability of amino acid R_i occurring at the subsite i ($-\xi_1, \ldots, -3, -2, -1, +1, +2, \ldots, +\xi_2$) for the secretion-cleavable segments, and $P_i^-(R_i)$ the corresponding probability for the non-secretion-cleavable segments. The values of the former can be derived from a positive training dataset S_0^+ consisting of only secretion-cleavable segments, and the values of the latter can be derived from a negative training dataset S_0^- consisting of only non-cleavable segments. The subscript 0 of Ψ indicates that the attribute function is formed by independent probabilities in which no coupling effect whatsoever among subsites is included, as shown by the right-hand side of eq. 2. However, in reality the protein subsites are often coupled with one another. If the coupling effect of a residue with those adjacent to it must be taken into account, then the probability factors in eqs.2a and 2b should be modified as

$$\Psi^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^+_{-\xi_1}(R_{-\xi_1})P^+_{-(\xi_1-1)}(R_{-(\xi_1-1)}|R_{-\xi_1}) \cdots P^+_{-2}(R_{-2}|R_{-3})P^+_{-1}(R_{-1}|R_{-2}) P^+_{+1}(R_{+1}|R_{-1})P^+_{+2}(R_{+2}|R_{+1}) \cdots P^+_{+\xi_2}(R_{+\xi_2}|R_{+(\xi_2-1)}) \quad (3a)$$

and

$$\Psi^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^-_{-\xi_1}(R_{-\xi_1})P^-_{-(\xi_1-1)}(R_{-(\xi_1-1)}|R_{-\xi_1}) \cdots P^-_{-2}(R_{-2}|R_{-3})P^-_{-1}(R_{-1}|R_{-2}) P^-_{+1}(R_{+1}|R_{-1})P^-_{+2}(R_{+2}|R_{+1}) \cdots P^-_{+\xi_2}(R_{+\xi_2}|R_{+(\xi_2-1)}) \quad (3b)$$

respectively, where $P^+_{-\xi_1}(R_{-\xi_1})$ and $P^-_{-\xi_1}(R-\xi_1)$ are the same as in eq. 2a and eq. 2b. $P^+_{-(\xi_1-1)}(R_{-(\xi_1-1)}|R_{-\xi_1})$ is the probability of amino acid $R_{-(\xi_1-1)}$ occurring at the subsite $-(\xi_1-1)$, given that $R_{-\xi_1}$ has occurred at the subsite $-\xi_1$; $P^+_{-2}(R_{-2}|R_{-3})$ is the probability of amino acid R_{-2} occurring at the subsite -2, given that R_{-3} has occurred at the subsite -3; and so forth. Their values can be derived from a positive training dataset S_0^+ consisting of only secretion-cleavable peptides. And $P^-_{-(\xi_1-1)}(R_{-(\xi_1-1)}|R_{-\xi_1})$, $P^-_{-2}(R_{-2}|R_{-3}), \ldots,$ have the same meaning as $P^+_{-(\xi_1-1)}(R_{-(\xi_1-1)}|R_{-\xi_1})$, $P^+_{-2}(R_{-2}|R_{-3}), \ldots,$ except that they are derived from a

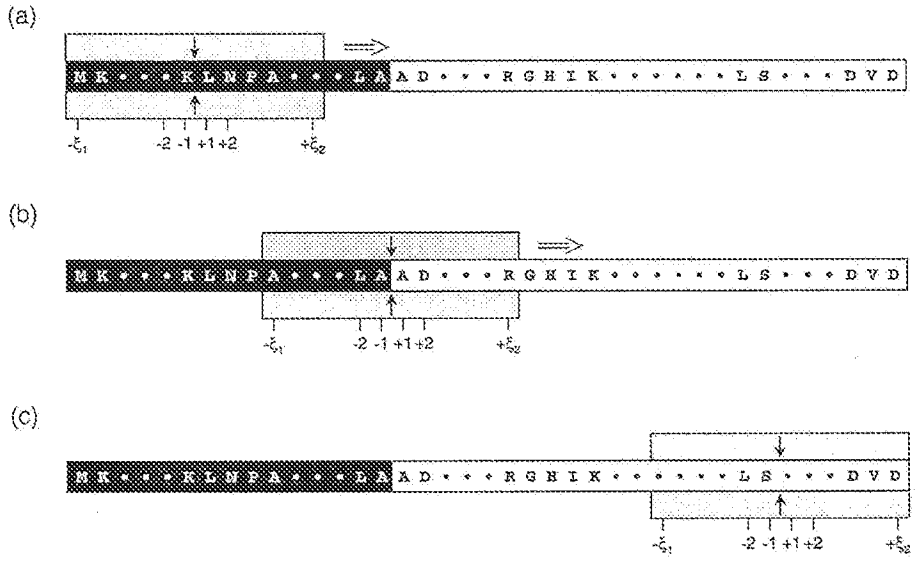


Fig. 3. Illustration of the scaled window model. When sliding the window $[-\xi_1, +\xi_2]$ along a protein sequence from the N-terminal (a) to C-terminal (c), the scales on the window are aligned with different amino acids so as to define different peptide segments. When, and only when, the scale −1 is aligned with the last residue of the signal sequence, and scale +1 aligned with the first residue of the mature protein as shown in panel (b), is the peptide segment seen within the window regarded as secretion-cleavable. Peptides segments seen within the window for all the other cases, such as those shown in panels (a) and (c), are regarded as non-secretion-cleavable. Amino acid residues in the signal part are expressed by white characters with black background white those in the mature protein by black characters with white background.

negative training dataset S_0^- consisting of only non-cleavable peptides.

Generally speaking, if the coupling effects of the μ (μ = 2, 3, . . .) closest neighboring amino acid residues need to be considered, then eq. 2 should be modified according to the μth-order Markov chain theory, i.e. the attribute function Ψ_0 should be replaced by Ψ_μ and the corresponding probability factors by the μth-order conditional probabilities. As one could surmise, the analysis of a higher-order Markov chain would be much more complicated. Therefore, the treatment in this paper is confined to the first-order Markov chain; i.e. only the first-order sequence-coupling effect is taken into account, as formulated by eq. 3.

Thus, for a given peptide segment as defined in eq. 1, if its attribute function to the positive training set S_0^+ is greater than that to the negative training set S_0^-, i.e. $\Psi^+ > \Psi^-$, then the sequence is predicted to be secretion-cleavable; otherwise, it is predicted to be a non-secretion-cleavable. We define a discriminant function Δ, given by

$$\begin{aligned}\Delta(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) \\ = w^+\Psi^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) \\ - w^-\Psi^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2})\end{aligned} \quad (4)$$

where w^+ and w^- are the weight factors for the attribute functions derived from the positive training dataset S_0^+ and negative training dataset S_0^-, respectively. If there is no special reason, they are generally set to be one; i.e. $w^+ = w^- = 1$. Thus, the criterion of the seretion-cleavable segment prediction for a given sequence can be formulated as follows:

$$\begin{cases}\text{The segment is secretion-cleavable,} & \text{if its } \Delta > 0 \\ \text{The segment is non-secretion-cleavable,} & \text{otherwise}\end{cases} \quad (5)$$

During the training process, the parameters ξ_1 and ξ_2 can be changed so as to find the optimal prediction quality. Once a secretion-cleavable segment is predicted, the corresponding cleavage site and signal peptide are automatically obtained as described above (cf. Fig. 3).

3. Results

To compare the power of different prediction algorithms, one must use a same data base; otherwise, the comparison would be meaningless. In view of this, we should adopt a dataset that is accessible to the public. The dataset investigated by Nielsen et al. [17] satisfies such a prerequisite; it can be retrieved from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. The dataset consists of 1939 secretory proteins and 1440 non-secretory proteins. The former contains 416 human, 1011 eukaryote, 105 *E.coli*, 266 Gram−, and 141 Gram+ proteins; while the latter 251 human, 820 eukaryote, 119 *E.coli*, 186 Gram−, and 64 Gram+ proteins. Redundant sequences were removed to guarantee that no pairs of homologous sequences exist in the data set. For the secretory proteins, the sequence of the signal peptide and the first 30 amino acids of the mature protein were included

in the datset, while for the non-secretory proteins, the first 70 amino acids of each sequence were included. Furthermore, to show the power of the current algorithm, the comparison should be made with the best result derived from the public-accessible dataset by the previous investigators. According to the report by Nielsen et al. [17], the average overall rate of correct prediction for the cleavage site was about 72%. As pointed out by Nielsen et al. [17], if "the original weight matrix algorithm [20] is applied to" their dataset, "the performance is much lower." Now let us apply the current prediction algorithm to the same dataset as used by Nielsen et al. [17], and see what results will be yielded.

The rates of correct prediction for the signal peptide set and non-signal peptide set are given by

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - m^+}{N^+}, & \text{for signal peptides} \\ \Lambda^- = \dfrac{N^- - m^-}{N^-}, & \text{for non-signal peptides} \end{cases} \tag{6}$$

where N^+ represents the total number of signal peptides, and m^+ is the number of signal peptides missed in prediction; N^- is the total number of non-signal peptides, and m^- is the number of non-signal peptides incorrectly predicted as signal peptides. The overall rate of correct prediction concerned is given by

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{m^+ + m^-}{N^+ + N^-} \tag{7}$$

We might have the situation of overprediction if the rate Λ^+ is very high but Λ^- very low; i.e. many non-signal peptides are incorrectly predicted as signal peptides. On the other hand, we might have the situation of underprediction if the rate Λ^+ is very low but Λ^- very high; i.e. many signal peptides are incorrectly predicted as non-signal peptides. Accordingly, the real prediction accuracy should be measured by Λ, the overall success rate. However, since the number of non-signal peptides is much greater than that of the signal peptides, we might also have the unpleasant situation where the overall rate Λ is very high but Λ^+ very low. The best compromise to solve this kind of situation is to keep Λ^+ with a decent rate while seeking for the highest rate for Λ. What rate for Λ^+ is decent? For the current case, it should be at least higher than 72%.

The prediction quality was examined by the standard testing procedure in statistics [15] that is a combination of the self-consistency and jackknife tests. In the former, the signal peptide of each protein in a given dataset was predicted using the parameters derived from the same dataset, the so-called training dataset; while in the latter, each protein in the training dataset was singled out in turn as a "test protein" and all the rule-parameters were derived from the remaining proteins. Compared with the independent dataset test and sub-sampling test often adopted in biology, the jackknife test is thought the most effective method for cross-validation in statistics [15]. This is because in the independent dataset test, the selection of a testing dataset is arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing dataset is sufficiently large [7]. As for the sub-sampling test in which a given dataset is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Nielsen et al. [17] each dataset was divided into five approximately equal size parts and then every network run was carried out with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different dataset divisions. Thus, according to their cross-validation scheme, the number of possible sub-sampling combinations would be, where $\Phi = \Phi^{\text{Sec}} \times \Phi^{\text{Non}}$, where $\Phi^{\text{Sec}} = \Phi^{\text{Sec}}_{\text{Hum}} \times \Phi^{\text{Sec}}_{\text{Euk}} \times \Phi^{\text{Sec}}_{\text{Eco}} \times \Phi^{\text{Sec}}_{\text{Gram}-} \times \Phi^{\text{Sec}}_{\text{Gram}+}$ is the number of possible sub-sampling combinations in the dataset of secretory proteins and $\Phi^{\text{Non}} = \Phi^{\text{Non}}_{\text{Hum}} \times \Phi^{\text{Non}}_{\text{Euk}} \times \Phi^{\text{Non}}_{\text{Eco}} \times \Phi^{\text{Non}}_{\text{Gram}-} \times \Phi^{\text{Non}}_{\text{Gram}+}$ the number of possible sub-sampling combinations in the dataset of non-secretory proteins. For the data studied by Nielsen et al. [17], we have $\Phi^{\text{Sec}}_{\text{Hum}} = 416!/(83!83!83!83!84!)$, $\Phi^{\text{Sec}}_{\text{Euk}} = 1011!/(202!202!202!202!203!)$, $\Phi^{\text{Sec}}_{\text{Eco}} = 105!/(21!21!21!21!21!)$, $\Phi^{\text{Sec}}_{\text{Gram}-} = 266!/(53!53!53!53!54!)$, $\Phi^{\text{Sec}}_{\text{Gram}+} = 141!/(28!28!28!28!29!)$, and $\Phi^{\text{Non}}_{\text{Hum}} = 251!/(50!50!50!50!51!)$, $\Phi^{\text{Non}}_{\text{Euk}} = 820!/(164!164!164!164!164!)$, $\Phi^{\text{Non}}_{\text{Eco}} = 119!/(24!24!24!24!23!)$, $\Phi^{\text{Non}}_{\text{Gram}-} = 186!/(37!37!37!37!38!)$, $\Phi^{\text{Non}}_{\text{Gram}+} = 64!/(13!13!13!13!12!)$. Of $\Phi^{\text{Sec}}_{\text{Hum}}$, $\Phi^{\text{Sec}}_{\text{Euk}}$, $\Phi^{\text{Sec}}_{\text{Eco}}$, $\Phi^{\text{Sec}}_{\text{Gram}-}$, and $\Phi^{\text{Sec}}_{\text{Gram}+}$, the smallest is $\Phi^{\text{Sec}}_{\text{Eco}} \sim 3.1 \times 10^{69}$, implying Φ^{Sec} would be $\geqslant 15.5 \times 10^{345}$. Of $\Phi^{\text{Non}}_{\text{Hum}}$, $\Phi^{\text{Non}}_{\text{Euk}}$, $\Phi^{\text{Non}}_{\text{Eco}}$, $\Phi^{\text{Non}}_{\text{Gram}-}$, and $\Phi^{\text{Non}}_{\text{Gram}+}$, the smallest is $\Phi^{\text{Non}}_{\text{Gram}+} \sim 1.76 \times 10^{41}$, implying Φ^{Non} would be $\geqslant 8.8 \times 10^{205}$. Thus, $\Phi = \Phi^{\text{Sec}} \times \Phi^{\text{Non}}$ would be $\geqslant 1.36 \times 10^{552}$. For such a huge number of combinations, it is impossible for any existing computer to handle. In fact in any practical sub-sampling tests as carried out by Nielsen et al. [17], only a very small fraction of the possible combinations were investigated, and the results thus obtained would unavoidably bear a considerable arbitrariness. Accordingly, the testing procedure adopted here is much more objective and rigorous.

Prediction was performed by selecting different parameters for the scaled window $[-\xi_1, +\xi_2]$. Preliminary jackknife tests indicated that for a given ξ_1 the optimal result for Λ^+ was obtained when $\xi_2 = 1$. Generally speaking, for the self-consistency test, the rate of correct prediction will be increased by widening the scaled window, i.e. increasing the values of ξ_1 (Table 1). However, for the jackknife test, increase of ξ_1 after its reaching a certain value will gradually reduce $-+$, the rate of correct prediction for signal peptides, although the overall rate of correct prediction keeps increasing (Table 1). Particularly, if ξ_1 is too large, many short signal peptides will be excluded for prediction.

Table 1
Performance values by using different parameters of the scaled window

Scaled window $[-\xi_1, +\xi_2]$	Success rate A^+ for signal peptide set[a]		Number of signal peptides excluded[b]	Overall success rate A^c	
	Self-consistency	Jackknife		Self-consistency	Jackknife
[−6, +1]	94.43%	79.78%	0	89.26%	89.20%
[−7, +1]	95.20%	79.42%	0	90.66%	90.58%
[−8, +1]	96.03%	78.34%	0	92.02%	91.88%
[−9, +1]	96.13%	78.49%	1	93.11%	92.90%
[−10, +1]	96.96%	78.34%	2	94.08%	93.84%
[−11, +1]	97.58%	77.20%	2	94.85%	94.60%
[−12, +1]	97.73%	76.48%	5	**95.46%**	**95.19%**
[−13, +1]	98.04%	75.81%	6	**95.99%**	**95.71%**
[−14, +1]	98.25%	74.98%	8	**96.25%**	**95.96%**
[−15, +1]	98.25%	73.49%	13	96.53%	96.21%
[−16, +1]	96.39%	69.98%	52	96.79%	96.44%

[a] This is the success rate for the 1939 signal peptide set; see eq. 6 for the definition of A^+.
[b] The excluded signal peptide should be counted as those missed in prediction, and hence is a part of m^+ as defined in eq. 6.
[c] See eq. 7 for the definition of the overall success rate A.

As shown in Table 1, when $\xi_1 \leq 8$, no signal peptides are excluded; when $\xi_1 = 9$, one signal peptide excluded; when $\xi_1 = 10$ or 11, two excluded; when $\xi_1 = 12$, five excluded; when $\xi_1 = 13$, six excluded; when $\xi_1 = 14$, eight; when $\xi_1 = 15$, thirteen; and so forth. These excluded signal peptides should also be counted as unsuccessful prediction events. To keep the number of excluded signal peptides being low and meanwhile to keep A^+ greater than 72%, we select $\xi_1 = 12$–14 and $\xi_2 = 1$. With these optimal parameters for the scaled window, the number of signal peptides excluded are less than 10 but the overall rates of correct prediction by both self-consistency and jackknife tests are over 95%.

4. Discussion

Since the current model is explicitly correlated with the sequential coupling along a peptide chain, it will provide a useful tool for helping further investigate many unclear molecular details regarding the molecular mechanism of the ZIP code protein-sorting system in cells, such as what will happen if an amino acid in the signal sequence is replaced by another, and how the signal sequence interacts with its counterpart of the signal peptidase. Moreover, the present method may also be used to improve the protein subcellular location prediction by establishing a more essential correlation of protein location directly with signal sequence, rather than the one indirectly with amino acid composition, as formulated in a number of papers in this area [3,6,8,9,19] and summarized in a recent review article [5]. It should be pointed out that the formulation presented here is a general one. By some modification, it can be used to study the coupling effects amoung some specific subsites as well [10]

Acknowledgments

Valuable discussions with Dr. A.P. Elhammer and Dr. Jinhe Li are gratefully acknowledged. The author would also like to thank Cindy Brennan, Raymond B. Moeller, Wendy Vanderheide and Katie Crawford of Pharmacia's Graphic Service Group for their help in drawing the figures.

References

[1] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Research 1997;25: 31–6.
[2] Bhat UN. Elements of Applied Stochastic Progresses, 2nd ed., Chap.3, New York: John Wiley & Sons, 1984. pp. 33–69.
[3] Cedano J, Aloy P, Perez-pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997; 266:594–600.
[4] Chou KC. Prediction of protein signal sequences and their cleavage sites. PROTEINS: Structure, Function, and Genetics 2000;42:136–9.
[5] Chou KC. Review: Prediction of protein structural classes and subcellular locations. Current Protein and Peptide Science 2000;1:171–208.
[6] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 2000; 278:477–83.
[7] Chou KC, Zhang CT. Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 1995;30: 275–349.
[8] Chou KC, Elrod DW. Protein subcellular location prediction. Protein Engineering 1999;12:107–18.
[9] Chou KC, Elrod DW. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 1998;252:63–8.
[10] Chou KC. Using subsite coupling to predict signal peptides. Protein Engineering 2001;14:75–9.
[11] Claros MG, Brunak S, von Heijne G. Prediction of N-terminal protein sorting signals. Curr Opin Struct Biol 1997;7:394–8.
[12] Gierasch LM. Signal sequences. Biochemistry 1989;28:923–30.
[13] Hagmann M. Colleagues say 'Amen' to this year's (Nobel Prizes) choices. Science 1999;286:666.

[14] King RD. In: Protein Structure Prediction: A Practical Approach, Sternberg MJE. (ed.), Oxford: IRL Press, 1996. pp. 79–97.
[15] Mardia KV, Kent JT, Bibby JM. Multivariate Analysis, 1979. p. 322 and p. 381, Academic Press, London.
[16] Nakai K. Protein sorting signals and prediction of subcellular localization. Advances in Protein Chemistry 2000;54:277–44.
[17] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 1997;10:1–6.
[18] Rapoport TA. Transport of proteins across the endoplasmic reticulum membrane. Science 1992;258:931–6.
[19] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–6.
[20] von Heijne G. A new method for predicting signal sequence cleavage sites. Nucleic Acids Research 1986;14:4683–90.
[21] Zheng N, Gierasch LM. Signal sequences: the same yet different. Cell 1996;86:849–52.

Protein Engineering vol.14 no.2 pp.75–79, 2001

COMMUNICATION

Using subsite coupling to predict signal peptides

Kuo-Chen Chou

Computer-Aided Drug Discovery, Pharmacia and Upjohn, Kalamazoo, MI 49007-4940, USA. E-mail: kuo-chen.chou@am.pnu.com

Given a nascent protein sequence, how can one predict its signal peptide or 'Zipcode' sequence? This is a first important problem for scientists to use signal peptides as a vehicle to find new drugs or to reprogram cells for gene therapy. Based on a model that takes into account the coupling effect among some key subsites, the so-called {–3, –1, +1} coupling model, a new prediction algorithm is developed. The overall rate of correct prediction for 1939 secretory proteins and 1440 non-secretary proteins was over 92%. It has not escaped our attention that the new method may also serve as a useful tool for helping investigate further many unclear details regarding the molecular mechanism of the ZIP code protein-sorting system in cells.

Keywords: {–3, –1, +1} coupling/non-secretory proteins/secretory proteins/'Zipcode' sequence

Introduction

The knowledge of protein signals can be used to reprogram cells in a specific way for future cell and gene therapy. Protein signals have become a crucial tool for researchers to construct new drugs that are targeted to a particular organelle to correct a specific defect. For example, by adding a specific tag to the desired proteins, one can tag them for excretion, making them much easier to harvest (Hagmann, 1999). To use such a tool successfully, first one has to identify the signal sequences. Since the number of nascent protein sequences entering databanks has been rapidly increasing, it is time consuming and costly to identify the signal peptides entirely by experiments. Thus, a strong interest in the automated identification of signal sequences and prediction of their cleavage sites has been evoked. The importance of predicting protein signal peptides has also been elaborated recently in an excellent review by Nakai (2000).

The existing methods in this area are based mostly on the use of neural networks (Claros *et al.*, 1997; Nielsen *et al.*, 1999; Nakai, 2000). They are actually the application of machine learning techniques. As pointed out by King (1996), the advantages of neural network prediction methods are that they are 'readily available' and 'often successful in practice'. He also pointed out that the disadvantages are that 'there is little use of chemical or physical theory', the methods have 'very poor explanatory power—a Hinton diagram means nothing to a protein chemist' and 'they are statistically rather poorly characterized'. Besides, although the computational costs for training the networks were considerably higher, the prediction accuracy thus obtained was not higher (and sometimes even lower) than the analytical methods. The current study was initiated in an attempt to develop an automated method based on the sub-site coupling principle that can be used to identify signal peptides faster and more accurately.

Materials and methods

Signal peptides comprise the N-terminal part of the secretory protein chain. They control the entry of virtually all proteins to the secretory pathway, in both eukaryotes and prokaryotes (Gierasch, 1989; Rapoport, 1992) and are cleaved off by signal peptidase while the protein is translocated through the membrane. As shown in Figure 1, the cleavage site is at (-1, +1), i.e. the location between residues –1 and +1 or between the last residue of the signal peptide and the first residue of the mature protein. Accordingly, the prediction of the signal peptide of a nascent protein is immediately correlated with the prediction of its cleavage site by the signal peptidase. The length of signal peptides is varied for different secretory proteins. As shown in Figure 2, of the 1939 signal peptides studied by Nielsen *et al.* (1997), the shortest one contains eight amino acid residues and the longest contains 90 residues while the majority have a length within 18–25 residues. The extreme variation in length and sequence has posed a difficulty for formulating a general algorithm to predict the signal peptides. To deal with this kind of situation, let us consider a window with a scale of ξ_1, ..., –3, –2, –1, +1, +2, ..., ξ_2 (Figure 3). Such a window is called a 'scaled window' and symbolized as $[-\xi_1, +\xi_2]$. When sliding the scaled window $[-\xi_1, +\xi_2]$ along a sequence of n residues, one can consecutively highlight $n - (\xi_1 + \xi_2) + 1$ different sequences. Note that for the current study the identification of cleavage site is very important because it is directly correlated with a correct prediction of the signal peptide. For example, instead of the site (-1, +1), if the cleavage site is identified at (–2, –1) or (+1, +2), then the corresponding signal peptide thus derived will be one residue shorter or longer than the actual one (Figure 1). Therefore, of the sequence segments highlighted by the scaled window, only the one with the residue at the scale –1 being the very last residue of the signal sequence is regarded as the secretion-cleavable segment (Figure 3a); while all the other segments regarded as non-secretion-cleavable (see, e.g., Figure 3b and c). In this way, if sliding the scaled window $[-\xi_1, +\xi_2]$ along a protein sequence of n residues, one can generate one, and only one, secretion-cleavable segment and $n - (\xi_1 + \xi_2)$ non-secretion-cleavable segments if the protein is secretory, but $n - (\xi_1 + \xi_2) + 1$ non-secretion-cleavable segments if it is non-secretory. All the secretion-cleavable segments form a cleavable or positive set denoted by S^+ and all the non-secretion-cleavable segments form a non-cleavable or negative set S^-.

Segments generated by sliding the scaled window $[-\xi_1, +\xi_2]$ along protein sequences can be generally expressed as

$$R_{-\xi_1}R_{-(\xi_1-1)}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+(\xi_2-1)}R_{+\xi_2} \qquad (1)$$

where $R_{-\xi_1}$ represents the residue at the scale $-\xi_1$, R_{-1} the residue at the scale –1, R_{+1} the residue at the scale +1 and so forth.

If the amino acid residue at each of the segment subsites (Equation 1) can be treated as an independent element, i.e.

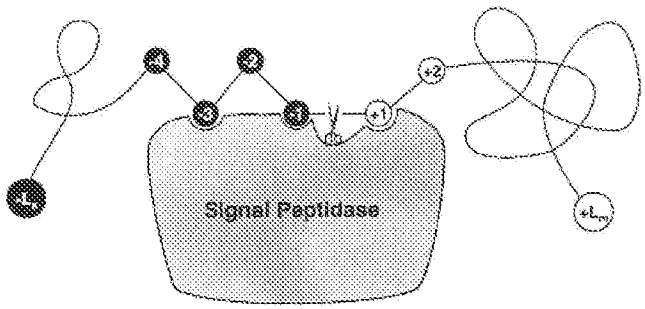


Fig. 1. A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a black circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with a black number. The cleavage site is at the position (–1, +1), i.e. between the last residue of the signal sequence and the first residue of the mature protein. During the cleavage process, a highly special fit is required between the amino acid residues at the subsites –3, –1 and +1 of the secretory protein and their counterpart of the enzyme (cf. Figure 4).

there is no coupling at all among these subsites, then its attribute to the cleavable set S^+ and that to the non-cleavable set S^- can be formulated, respectively, as

$$\Psi_0^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^+_{-\xi_1}(R_{-\xi_1})\cdots P^+_{-3}(R_{-3})P^+_{-2}(R_{-1}) P^+_{+1}(R_{+1})P^+_{+2}(R_{+2})\cdots P^+_{+\xi_2}(R_{+\xi_2}) \quad (2a)$$

and

$$\Psi_0^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^-_{-\xi_1}(R_{-\xi_1})\cdots P^-_{-3}(R_{-3})P^-_{-2}(R_{-1}) P^-_{+1}(R_{+1})P^-_{+2}(R_{+2})\cdots P^-_{+\xi_2}(R_{+\xi_2}) \quad (2b)$$

where $P_i^+(R_i)$ is the probability of amino acid R_i occurring at the subsite i ($= -\xi_1, \ldots, -3, -2, -1, +1, +2, \ldots, +\xi_2$) for the secretion-cleavable segments and $P_i^-(R_i)$ the corresponding probability for the non-secretion-cleavable segments. The values of the former can be derived from a positive training data set S_0^+ consisting of only secretion-cleavable segments and the values of the latter can be derived from a negative training data set S_0^- consisting of only non-cleavable segments. The subscript 0 of ψ indicates that the attribute function is formed by independent probabilities in which no coupling effect between subsites is included, as shown by the right-hand side of Equation 2. However, in reality the protein subsites are often coupled with one another. Therefore, it is instructive to conduct a statistical analysis for the 1939 secretory protein sequences retrieved from Nielsen *et al.* (1997). The result thus obtained is illustrated in Figure 4, from which we can see that the amino acid residues at the subsites –3, –1 and +1 are mostly occupied by Ala. Furthermore, according to the detailed numbers generated through the statistical analysis, of the 1939 protein sequences, the occurrence frequencies of Ala at the subsites –3, –1 and +1 are 667, 1084 and 397, respectively, while the occurrence frequencies of the other 19 amino acids at these subsites are relatively much lower. Besides, all these three subsites are very close to the cleavage site (Figure 1). This suggests that a highly special match between the signal peptidase and the secretory protein at the subsites –3, –1 and +1 is required during the cleavage process. Accordingly, to establish a powerful method for predicting the signal peptides, the coupling among these three key subsites, i.e. the {–3, –1, +1} coupling, must be taken into account. Thus, Equations 2a and 2b should be modified to

$$\Psi^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^+_{-\xi_1}(R_{-\xi_1})\cdots P^+_{-3}(R_{-3})P^+_{-2}(R_{-2})P^+_{-1}(R_{-1}|R_{-3}) P^+_{+1}(R_{+1}|R_{-1})P^+_{+2}(R_{+2})\cdots P^+_{+\xi_2}(R_{+\xi_2}) \quad (3a)$$

and

$$\Psi^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = P^-_{-\xi_1}(R_{-\xi_1})\cdots P^-_{-3}(R_{-3})P^-_{-2}(R_{-2})P^-_{-1}(R_{-1}|R_{-3}) P^-_{+1}(R_{+1}|R_{-1})P^-_{+2}(R_{+2})\cdots P^-_{+\xi_2}(R_{+\xi_2}) \quad (3b)$$

respectively, where $P_i^+(R_i)$ and $P_i^-(R_i)$ are the same as those in Equation 2. $P^+_{-1}(R_{-1}|R_{-3})$ is the probability of amino acid R_{-1} occurring at the subsite –1, given that R_{-3} has occurred at the subsite –3; $P^+_{+1}(R_{+1}|R_{-1})$ is the probability of amino acid R_{+1} occurring at the subsite +1, given that R_{-1} has occurred at the subsite –1. Their values can be derived from a positive training data set S_0^+ consisting of only secretion-cleavable peptides. Also, $P^-_{-1}(R_{-1}|R_{-3})$ and $P^-_{+1}(R_{+1}|R_{-1})$ have the same meaning as $P^+_{-1}(R_{-1}|R_{-3})$ and $P^+_{+1}(R_{+1}|R_{-1})$ except that they are derived from a negative training data set S_0^- consisting of only non-cleavable peptides.

Thus, for a given peptide sequence as defined in Equation 1, if its attribute function to the positive training set S_0^+ is greater than that to the negative training set S_0^-, i.e. $\psi^+ > \psi^-$, then the sequence is predicted to be secretion-cleavable; otherwise, it is predicted to be non-secretion-cleavable. We define a discriminant function Δ, given by

$$\Delta(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) = w^+\Psi^+(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2}) - w^-\Psi^-(R_{-\xi_1}\cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+\xi_2} \quad (4)$$

where w^+ and w^- are the weight factors for the attribute functions derived from the positive training data set S_0^+ and negative training data set S_0^-, respectively. If there is no special reason, they are generally set to be one i.e. $w^+ = w^- = 1$. Thus, the criterion of predicting the seretion-cleavability for a given peptide sequence can be formulated as follows:

$$\begin{cases} \text{The peptide is secretion-cleavable,} & \text{if its } \Delta > 0 \\ \text{The peptide is non-secretion-cleavable,} & \text{otherwise} \end{cases} \quad (5)$$

During the training process, the parameters ξ_1 and ξ_2 can be changed so as to find the optimal prediction quality. Once a secretion-cleavable peptide is predicted, the corresponding cleavage site and signal peptide are automatically obtained as described above (cf. Figures 1 and 3a).

Results and discussion

To show the power of the key-subsites-coupled algorithm, the following two criteria should be followed: (1) using a good data set that is accessible to the public and (2) comparison with the best result reported in the literature. The data set investigated by Nielsen *et al.* (1997) satisfies the first criterion; it can be retrieved from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. They consist of 1939 secretory proteins and 1440

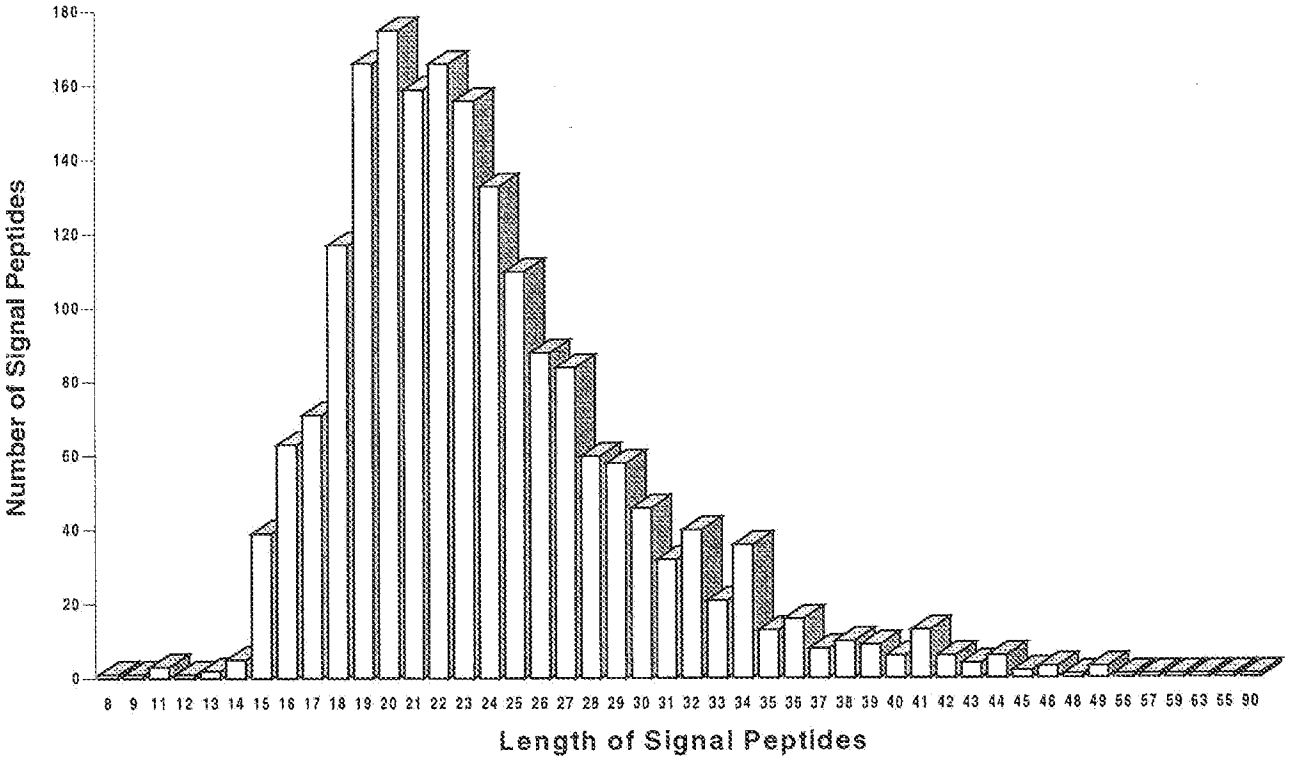


Fig. 2. A histogram to show the distribution of signal peptides with their length in the 1939 secretory proteins retrieved from Nielsen *et al.* (1997).

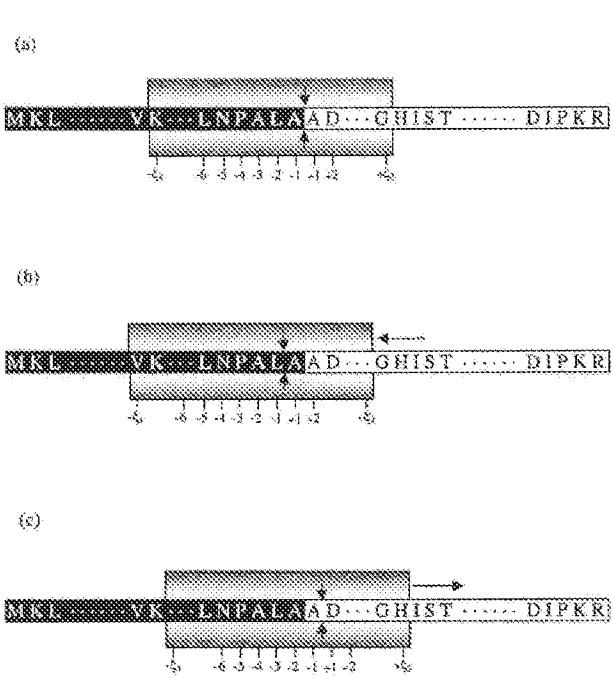


Fig. 3. Illustration to show the sequence segments highlighted by sliding the scaled window $[-\xi_1, +\xi_2]$ along a protein sequence. During the sliding process, the scales on the window are aligned with different amino acids so as to define different peptide segments. When, and only when, the scale −1 is aligned with the last residue of the signal sequence and scale +1 aligned with the first residue of the mature protein as shown in panel (a) is the peptide segment seen within the window regarded as secretion-cleavable. Peptides segments seen within the window for all the other cases, such as those shown in panels (b) and (c), are regarded as non-secretion-cleavable.

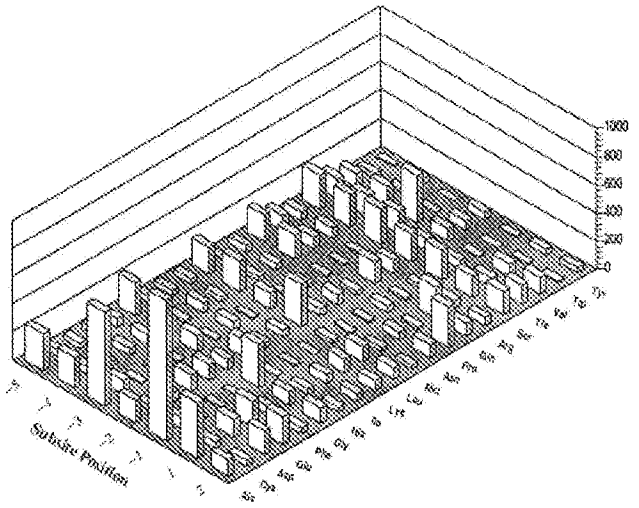


Fig. 4. A 3-D histogram to show the frequencies of the 20 native amino acids that occur at the subsites proximal to the cleavage site. As shown, the occurrence frequencies of Ala at the subsites −3, −1 and +1 are overwhelming in comparison with the other 19 amino acids, suggesting a high selectivity of Ala at the three key subsites (cf. Figure 1).

non-secretory proteins. The former contains 416 human, 1011 eukaryote, 105 *Escherichia coli*, 266 Gram negative and 141 Gram positive proteins, and the latter 251 human, 820 eukaryote, 119 *E.coli*, 186 Gram negative and 64 Gram positive proteins. Redundant sequences were removed to guarantee that no pairs of homologous sequences exist in the data set. As treated by Nielsen *et al.* (1997), for the secretory proteins, the sequence of the signal peptide and the first 30 amino acids of the mature protein were included in the data set, whereas for

the non-secretory proteins, the first 70 amino acids of each sequence were included. According to their report, the average rate of correct prediction for the cleavage site location by the neural network method was 71.54%. This is the highest success rate so far reported for such a large data set available to the public. Therefore, the result reported by Nielsen *et al.* (1997) also satisfies the second criterion. To compare the prediction quality at an equivalent condition, we used the same data set as used by Nielsen *et al.* (1997).

The rate of correct prediction for the signal peptide set and non-signal peptide set are given by

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - m^+}{N^+}, & \text{for signal peptides} \\ \Lambda^- = \dfrac{N^- - m^-}{N^-}, & \text{for non-signal peptides} \end{cases} \quad (6)$$

where N^+ represents the total number of signal peptides and m^+ is the number of signal peptides missed in prediction; N^- is the total number of non-signal peptides and m^- is the number of non-signal peptides incorrectly predicted as signal peptide. The overall rate of correct prediction concerned is given by

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{m^+ + m^-}{N^+ + N^-} \quad (7)$$

The prediction quality was examined by the standard testing procedure in statistics (Mardia *et al.*, 1979), that is, a combination of the self-consistency and jackknife tests. In the former, the signal peptide of each protein in a given data set was predicted using the parameters derived from the same data set, the so-called training data set, whereas in the latter, each protein in the training data set was singled out in turn as a 'test protein' and all the rule-parameters were derived from the remaining proteins. Compared with the independent data set test and sub-sampling test often adopted in biology, the jackknife test is considered to be the most effective method for cross-validation in statistics (Mardia *et al.*, 1979). This is because in the independent data set test, the selection of a testing data set is arbitrary and the accuracy thus obtained lacks an objective criterion unless the testing data set is sufficiently large (Chou and Zhang, 1995). As for the sub-sampling test in which a given data set is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Nielsen *et al.* (1977), each data set was divided into five approximately equal size parts and then every network run was carried out with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different data set divisions. Thus, even for the data of only secretory proteins, the number of possible combinations would be $\Phi = \Phi_1 \times \Phi_2 \times \Phi_3 \times \Phi_4 \times \Phi_5$, where $\Phi_1 = 416!/(83!83!83!83!84!)$, $\Phi_2 = 1011!/(202!202!202!202!202!)$, $\Phi_3 = 105!/(21!21!21!21!21!)$, $\Phi_4 = 266!/(53!53!53!53!54!)$ and $\Phi_5 = 141!/(28!28!28!28!29!)$. Of Φ_1, Φ_2, Φ_3, Φ_4 and Φ_5, the smallest is $\Phi_3 \approx 3.1 \times 10^{69}$, implying Φ would be $\gg 15.5 \times 10^{345}$. It is impossible for any existing computer to handle such a huge number of combinations. In fact in any practical sub-sampling tests as performed by Nielsen *et al.* (1997), only a very small fraction of the possible combinations were investigated and the results thus obtained could not avoid a considerable arbitrariness. Accordingly, the testing procedure adopted here is much more objective and rigorous.

Table I. Performance values by using the subsite coupling model

	Rate of correct prediction for cleavage site location (%)[a]		
	Signal peptides	Non-secretory proteins	Overall
Scaled window	Self-consistency test		
$[-\xi_1, +\xi_2]$	Λ^+	Λ^-	Λ
[–6, +2]	89.84	87.44	87.47
[–8, +2]	90.36	89.10	89.12
[–10, +2]	92.06	90.76	90.78
[–12, +2]	93.66	92.11	92.13
[–13, +2]	93.96	92.46	92.48
[–14, +2]	93.97	92.57	92.59
[–15, +2]	93.97	92.67	92.69
[–16, +2]	92.26	92.75	92.74
[–18, +2]	86.02	93.09	92.99
Scaled window	Jackknife test		
$[-\xi_1, +\xi_2]$	Λ^+	Λ^-	Λ
[–6, +2]	85.25	87.69	87.66
[–8, +2]	86.90	89.15	89.12
[–10, +2]	87.98	90.74	90.71
[–12, +2]	89.12	92.10	92.06
[–13, +2]	89.63	92.46	92.42
[–14, +2]	89.58	92.57	92.53
[–15, +2]	89.94	92.66	92.63
[–16, +2]	88.14	92.74	92.68
[–18, +2]	81.74	93.08	92.93

[a]See Equations 6 and 7 for the definitions of Λ^+, Λ^- and Λ.

Prediction was performed by selecting different parameters for the scaled window $[-\xi_1, +\xi_2]$. Preliminary tests indicated that for a given ξ_1 the optimal result for Λ^+ was obtained when $\xi_2 = 2$. The predicted results by both self-consistency and jackknife tests with different values of ξ_1 are given in Table I, from which we can see that the overall success rate Λ is improved with increase in ξ_1. However, if ξ_1 is too large, many short signal peptides will be excluded. For example, two signal peptides were excluded when $\xi_1 = 10$, five when $\xi_1 = 12$, six when $\xi_1 = 13$, eight when $\xi_1 = 14$, 13 when $\xi_1 = 15$, 52 when $\xi_1 = 16$ and 186 when $\xi_1 = 18$. Each of these excluded signal peptides was counted as an unsuccessful prediction event, contributing to the reduction of the success rate for the prediction of signal peptides. As a consequence, Λ^+ was gradually reduced when $\xi_1 \geqslant 16$ (Table I). As a compromise, we select $\xi_1 = 13$, 14 or 15 and $\xi_2 = 2$ as the optimal parameters for the scaled window $[-\xi_1, +\xi_2]$. When ξ_1 and ξ_2 are within these values, the success rates Λ^+ (Equation 6) for the signal peptide set are over 93 and 89% by self-consistency and jackknife tests, respectively, while the corresponding success rates Λ^- (Equation 6) for the non-signal peptide set are both over 92%. Also, the overall success rates (Equation 7) for the cleavage site location by both self-consistency and jackknife tests are over 92%.

Besides the neural network (NN) method proposed by Nielsen *et al.* (1997), there are some other methods, such as the simple weight matrix method (von Heijne, 1986), the hidden Markov method (Baldi and Brunak, 1998) and the physical sequence analysis method (Ladunga, 1999). Like Nielsen *et al.*'s method, all these methods have played an

important role in stimulating the development of this area. The simple weight matrix method is one of the earliest practical approaches for predicting the signal peptide cleavage sites. However, as pointed out by Nielsen *et al.* (1997), if 'the original weight matrix algorithm (von Heijne, 1986) is applied to' the current data set, 'the performance is much lower' in comparison with their NN method. The hidden Markov method (HMM) also belongs to the machine learning approach; the term 'hidden' refers to the invisibility of the underlying random walk between different states. Actually, the HMM method is a different type of artificial neural network method and hence also bears the disadvantages elaborated by King (1996). The physical sequence analysis method, also called PHYSEAN method, was established on the basis of the physical, chemical and biological characteristics of protein sequences. The working data sets for PHYSEAN consists of 2532 preproteins with signal peptides and 1138 cytosolic proteins. As described by Ladunga (1999), three-quarters of the sequences in the data sets were randomly selected to form a training set and the predictions were performed on the remaining one-quarter of sequences. The prediction accuracy was estimated on untrained proteins by five repetitions of cross-validation experiments. The success rate thus obtained for the prediction of cleavage sites was 79.28%. It was not possible to make a direct comparison of the present algorithm with PHYSEAN based on a same data set because, unlike NN (Nielsen *et al.*, 1997), the data sets in PHYSEAN are not accessible to the public. Moreover, as we can see, the cross-validation procedure in PHYSEAN is also of sub-sampling test and hence could not avoid the problem of arbitrariness either. This can be illustrated as follows. Even only for the 2532 preproteins, the number of possible sub-sampling combinations would be $2532!/(633!1899!) \gg 10^{370}$. Compared with such a huge number, five different sub-samplings, although randomly selected, are merely a very tiny fraction of the possible combinations (i.e., the fraction of sub-samplings considered is $\ll 0.5\times10^{-369}$).

Accordingly, from both the higher success rate and the more rationality in test procedures, it is worth communicating the new algorithm to those working in the area concerned. At least it will play a complementary role to the existing algorithms, stimulating the development of protein signal peptide prediction.

Conclusion

Since the present model has explicitly incorporated the coupling among the subsites −3, −1 and +1 and all these subsites are very close to the cleavage site, it can be directly used for investigating the protein secretion-cleaved mechanism by signal peptidase. The present model can also serve as a useful vehicle for helping further investigate many unclear details regarding the molecular mechanism of the ZIP code protein-sorting system in cells. Furthermore, since signal peptides are the key in determining the subcellular location of proteins, the {−3, −1, +1} model might have some impact in improving the prediction quality of protein subcellular location (Cedano *et al.*, 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1998, 1999a,b; Chou, 2000; Nakai, 2000).

Acknowledgements

Illuminative discussions with Dr A.P.Elhammer and Dr Jinhe Li are gratefully acknowledged. The author also thanks Raymond B.Moeller, Cynthia A.Ludlow and Wendy Vanderheide for their help in drawing the figures.

References

Baldi,P. and Brunak,S. (1998) *Bioinformatics: the Machine Learning Approach*. MIT Press, Cambridge, MA.
Cedano,J., Aloy,P., Perez-Pons,J.A. and Querol,E. (1997) *J. Mol. Biol.*, **266**, 594–600.
Chou,K.C. (2000) *Curr. Protein Pept. Sci.*, **1**, 171–208.
Chou,K.C. and Elrod,D.W. (1998) *Biochem. Biophys. Res. Commun.*, **252**, 63–68.
Chou,K.C. and Elrod,D.W. (1999a) *Protein Eng.*, **12**, 107–118.
Chou,K.C. and Elrod,D.W. (1999b) *Proteins: Struct. Funct. Genet.*, **34**, 137–153.
Chou,K.C. and Zhang,C.T. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.
Claros,M.G., Brunak,S. and von Heijne,G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 394–398.
Gierasch,L.M. (1989) *Biochemistry*, **28**, 923–930.
Hagmann,M. (1999) *Science*, **286**, 666–666.
King,R.D. (1996) In Sternberg,M.J.E. (ed.), *Protein Structure Prediction: a Practical Approach*. IRL Press, Oxford, pp. 79–97.
Ladunga,I. (1999) *Bioinformatics*, **15**, 1028–1038.
Mardia,K.V., Kent,J.T. and Bibby,J.M. *Multivariate Analysis*. Academic Press, London, 1979, pp. 322 and 381.
Nakai,K. (2000) *Adv. Protein Chem.*, **54**, 277–344.
Nielsen,H., Engelbrecht,J., Brunak S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.
Nielsen,H., Brunak S. and von Heijne,G. (1999) *Protein Eng.*, **12**, 3–9.
Rapoport,T.A. (1992) *Science*, **258**, 931–936.
Reinhardt,A. and Hubbard,T. (1998) *Nucleic Acids Res.*, **26**, 2230–2236.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

Received October 13, 2000; revised November 29, 2000; accepted December 8, 2000

Protein Engineering vol.10 no.1 pp.1–6, 1997

SHORT COMMUNICATION

Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites

Henrik Nielsen, Jacob Engelbrecht[1], Søren Brunak and Gunnar von Heijne[2]

Center for Biological Sequence Analysis, Department of Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark and [2]Department of Biochemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

[1]Present address: Novo Nordisk A/S, Scientific Computing, Building 9M1, Novo Allé, DK-2880 Bagsværd, Denmark

We have developed a new method for the identification of signal peptides and their cleavage sites based on neural networks trained on separate sets of prokaryotic and eukaryotic sequence. The method performs significantly better than previous prediction schemes and can easily be applied on genome-wide data sets. Discrimination between cleaved signal peptides and uncleaved N-terminal signal-anchor sequences is also possible, though with lower precision. Predictions can be made on a publicly available WWW server.

Keywords: cleavage sites/protein sorting/secretion/signal peptide

Introduction

Signal peptides control the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes (Gierasch, 1989; von Heijne, 1990; Rapoport, 1992). They comprise the N-terminal part of the amino acid chain and are cleaved off while the protein is translocated through the membrane. The common structure of signal peptides from various proteins is commonly described as a positively charged n-region, followed by a hydrophobic h-region and a neutral but polar c-region. The (–3,–1) rule states that the residues at positions −3 and −1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly (von Heijne, 1983, 1985).

A strong interest in the automated identification of signal peptides and the prediction of their cleavage sites has been evoked not only by the huge amount of unprocessed data available, but also by the industrial need to find more effective vehicles for the production of proteins in recombinant systems. The most widely used method for predicting the location of the cleavage site is a weight matrix which was published in 1986 (von Heijne, 1986). This method is also useful for discriminating between signal peptides and non-signal peptides by using the maximum cleavage site score. The original matrices are commonly used today, even though the amount of signal peptide data available has increased since 1986 by a factor of 5–10.

Here, we present a combined neural network approach to the recognition of signal peptides and their cleavage sites, using one network to recognize the cleavage site and another network to distinguish between signal peptides and non-signal peptides. A similar combination of two pairs of networks has been used with success to predict the intron splice sites in pre-mRNA from humans and the dicotelydoneous plant *Arabidopsis thaliana* (Brunak *et al.*, 1991; S.Hebsgaard, P.Korning, J.Engelbrecht, P.Rouzé and S.Brunak, submitted). Artificial neural networks have been used for many biological sequence analysis problems (Hirst and Sternberg, 1992; Presnell and Cohen, 1993). They have also been applied to the twin problems of predicting signal peptides and their cleavage sites, but until now without leading to practically applicable prediction methods with significant improvements in performance compared with the weight matrix method (Arrigo *et al.*, 1991; Ladunga *et al.*, 1991; Schneider and Wrede, 1993).

Materials and methods

The data were taken from SWISS-PROT version 29 (Bairoch and Boeckmann, 1994). The data sets were divided into prokaryotic and eukaryotic entries and the prokaryotic data sets were further divided into Gram-positive eubacteria (*Firmicutes*) and Gram-negative eubacteria (*Gracilicutes*), excluding *Mycoplasma* and *Archaebacteria*. Viral, phage and organellar proteins were not included. In addition, two single-species data sets were selected, a human subset of the eukaryotic data and an *Escherichia coli* subset of the Gram-negative data.

The sequence of the signal peptide and the first 30 amino acids of the mature protein from the secretory protein were included in the data set. The first 70 amino acids of each sequence were used from the cytoplasmic and (for the eukaryotes) nuclear proteins. In addition, a set of eukaryotic signal anchor sequences, i.e. N-terminal parts of type II membrane proteins (von Heijne, 1988), were extracted (see Figure 1).

As an example of a large-scale application of the finished method, we used the *Haemophilus influenzae* Rd genome—the first genome of a free-living organism to be completed (Fleischmann *et al.*, 1995). We have downloaded the sequences of all the predicted coding regions in the *H.influenzae* genome from the World Wide Web (WWW) server of the Institute for Genomic Research at http://www.tigr.org/. Only the first 60 positions of each sequence were analysed.

We have attempted to avoid signal peptides where the cleavage sites are not experimentally determined, but we are not able to eliminate them completely, since many database entries simply lack information about the quality of the evidence. The details of the data selection are described in the WWW server and in an earlier paper (Nielsen *et al.*, 1996a).

Redundancy in the data sets was avoided by excluding pairs of sequences which were functionally homologous, i.e. those that had more than 17 (eukaryotes) or 21 (prokaryotes) exact matches in a local alignment (Nielsen *et al.*, 1996a). Redundant sequences were removed using an algorithm which guarantees that no pairs of homologous sequences remain in the data set (Hobohm *et al.*, 1992). This procedure removed 13–56% of the sequences. The numbers of non-homologous sequences remaining in the data sets are shown in Table I. Redundancy

Table I. Data and performance values

Source	Data		Network architecture (window/hidden units)		Performance	
	(Number of sequences)					
	Signal peptides	Non-secretory proteins	*C*-score	*S*-score	Cleavage site location (% correct)	Signal peptide discrimination (correlation)
Human	416	251	15+4/2	27 / 4	68.0 (67.9)	0.96 (0.97)
Eukaryote	1011	820	17+2/2	27 / 4	70.2	0.97
E.coli	105	119	15+2/2	39 / 0	83.7 (85.7)	0.89 (0.92)
Gram–	266	186	11+2/2	19 / 3	79.3	0.88
Gram+	141	64	21+2/0	19 / 3	67.9	0.96

Data: the number of sequences of signal peptides and non-secretory (i.e. cytoplasmic or nuclear) proteins in the data sets after redundancy reduction. The organism groups are eukaryotes, human, Gram-negative bacteria ('Gram–'), *E.coli* and Gram-positive bacteria ('Gram+'). The human data are subsets of the eukaryotic data and the *E.coli* data are subsets of the Gram-negative data. The signal anchor and *H.influenzae* data are not shown in the table. *Network architecture*: the size of the input window and the number of hidden computational units ('neurons') in the optimal neural networks chosen for each data set. *C*-score networks have asymmetrical input windows. *Performance*: the percentage of signal peptide sequences where the cleavage site was predicted to be at the correct location according to the maximal value of the Y-score (see Figure 2). The ability of the method to distinguish between the signal peptides and the N-terminals of non-secretory proteins (based on the mean value of the *S*-score in the region between position 1 and the predicted cleavage site position) is measured by the correlation coefficients (Mathews, 1975). Both performance values are measured on the test sets (the average of five cross-validation tests). The values given in parentheses indicate the performance for the human sequences when using networks trained on all eukaryotic data and for the *E.coli* sequences when using Gram-negative networks respectively.

reduction was not applied to the signal anchor data or the *H.influenzae* data, since these were not used as training data.

Neural network algorithms

The signal peptide problem was posed to the neural networks in two ways: (i) recognition of the cleavage sites against the background of all other sequence positions and (ii) classification of amino acids as belonging to the signal peptide or not. In the latter case, negative examples included both the first 70 positions of non-secretory proteins and the first 30 positions of the mature part of secretory proteins.

The neural networks were feed-forward networks with zero or one layer of two to 10 hidden units, trained using back-propagation (Rumelhart *et al.*, 1986) with a slightly modified error function. The sequence data were presented to the network using sparsely encoded moving windows (Qian and Sejnowski, 1988; Brunak *et al.*, 1991). Symmetric and asymmetric windows of a size varying from five to 39 positions were tested.

Based on the numbers of correctly and incorrectly predicted positive and negative examples, we calculated the correlation coefficient (Mathews, 1975). The correlation coefficients of both the training and test sets were monitored during training and the performance of the training cycle with the maximal test set correlation was recorded for each training run. The networks chosen for inclusion in the WWW server have been trained until this cycle only.

The test performances have been calculated by cross-validation: each data set was divided into five approximately equal-sized parts and then every network run was carried out with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different data set divisions.

For each of the five data sets, one signal peptide/non-signal peptide network architecture and one cleavage site/non-cleavage site network architecture was chosen on the basis of the test set correlation coefficients. We did not pick the architecture with absolutely the best performance, but instead the smallest network that could not be significantly improved by enlarging the input window or adding more hidden units.

The trained networks provide two different scores between zero and one for each position in an amino acid sequence. The output from the signal peptide/non-signal peptide networks, the *S*-score, can be interpreted as an estimate of the probability of the position belonging to the signal peptide, while the output from the cleavage site/non-cleavage site networks, the *C*-score, can be interpreted as an estimate of the probability of the position being the first in the mature protein (position +1 relative to the cleavage site).

If there are several *C*-score peaks of comparable strength, the true cleavage site may often be found by inspecting the *S*-score curve in order to see which of the *C*-score peaks coincides best with the transition from the signal peptide to the non-signal peptide region. In order to formalize this and improve the prediction, we have tried a number of linear and non-linear combinations of the raw network scores and evaluated the percentage of sequences with correctly placed cleavage sites in the five test sets. The best measure was the geometric average of the C-score and a smoothed derivative of the S-score, termed the Y-score:

$$Y_i = \sqrt{C_i \Delta_d S_i} \qquad (1)$$

where $\Delta_d S_i$ is the difference between the average *S*-score of d positions before and d position after position i:

$$\Delta_d S_i = \frac{1}{d}\left(\sum_{j=1}^{d} S_{i-j} - \sum_{j=0}^{d-1} S_{i+j} \right) \qquad (2)$$

In Figure 2(A), examples of the values of the C-, S- and Y-scores are shown for a typical signal peptide with a typical cleavage site. The C-score has one sharp peak that corresponds to an abrupt change in the S-score from a high to low value. Among the real examples, the C-score may exhibit several peaks and the S-score may fluctuate. We define a cleavage site as being correctly located if the true cleavage site position corresponds to the maximal Y-score (combined score).

For a typical non-secretory position, the values of the C-, S- and Y-scores are lower, as shown in Figure 2(B). We found the best discriminator between signal peptides and non-secretory

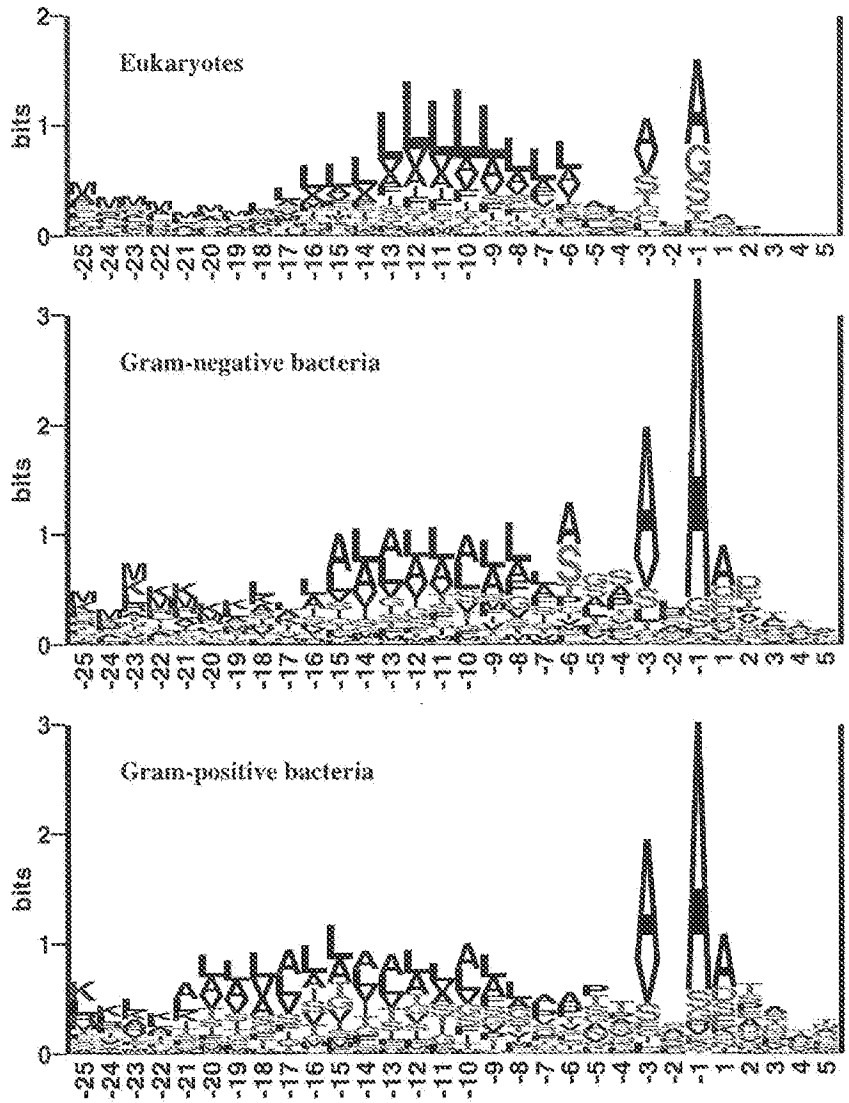


Fig. 1. Sequence logos (Schneider and Stephens, 1990) of signal peptides, aligned by their cleavage sites. The total height of the stack of letters at each position shows the amount of information, while the relative height of each letter shows the relative abundance of the corresponding amino acid. The information is defined as the difference between the maximal and actual entropy (Shannon, 1948): $I_j = H_{max} - H_j = \log_2 20 + \Sigma_\alpha\, n_j(\alpha)/N_j \log_2 n_j(\alpha)/N_j$, where $n_j(\alpha)$ is the number of occurrences of the amino acid α and N_j is the total number of letters (occupied positions) at position j. Positively and negatively charged residues are shown in blue and red respectively, while uncharged polar residues are green and hydrophobic residues are black.

proteins to be the average of the S-score in the predicted signal peptide region, i.e. from position 1 to the position immediately before the position where the Y-score has a maximal value. If this value—the mean S-score—is greater than 0.5, we predict the sequence in question to be a signal peptide (cf. Figure 3).

The relationship between the various performance measures and their development during the training process is described in detail elsewhere (Nielsen *et al.*, 1997).

Results and discussion

The optimal network architecture and corresponding predictive performance for all the data sets are shown in Table I. The C-score problem is best solved by networks with asymmetric windows, i.e. windows including more positions upstream than downstream of the cleavage site. This corresponds well with the location of the cleavage site pattern information which is shown as sequence logos (Schneider and Stephens, 1990) in Figure 1. The S-score problem, on the other hand, is best solved by symmetric or approximately symmetric windows.

Although our method is able to locate cleavage sites and discriminate signal peptides from non-secretory proteins with a reasonably high reliability, the accuracy of the cleavage site location is lower than that reported for the original weight matrix method (von Heijne, 1986): 78% for eukaryotes and

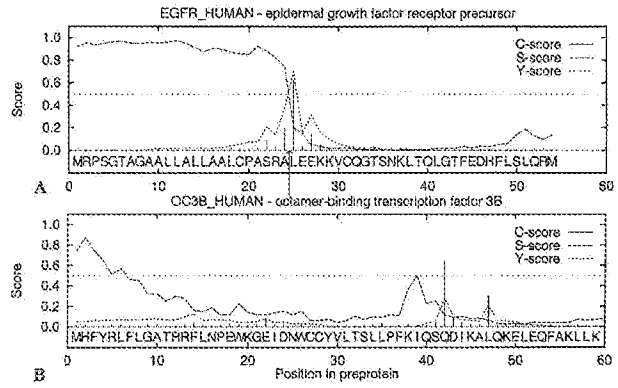


Fig. 2. Examples of network output. The values of the C- (output from cleavage site networks), S- (output from signal peptide networks) and Y-scores (combined cleavage site score, $Y_k = \sqrt{C_k \Delta_k S_k}$) are shown for each position in the sequence. The C- and S-scores are averages over five networks trained on different parts of the data. Note: the C- and Y-scores are high for the position immediately after the cleavage site, i.e. the first position in the mature protein. (A) A successfully predicted signal peptide. The true cleavage site is marked with an arrow. (B) A non-secretory protein. For many non-secretory proteins, all three scores are very low throughout the sequence. In this example, there are peaks of the C- and S-scores, but the sequence is still easily classified as non-secretory, since the C-score peak occurs far away from the S-score decline and the region of the high S-score is far too short.

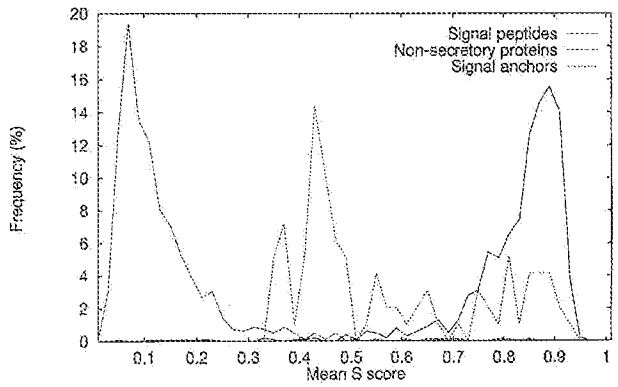


Fig. 3. Distribution of the mean signal peptide score (S-score) for signal peptides and non-signal peptides (eukaryotic data only). 'Non-secretory proteins' refer to the N-terminal parts of cytoplasmic or nuclear proteins, while 'signal anchors' are the N-terminal parts of type II membrane proteins. The mean S-score of a sequence is the average of the S-score over all positions in the predicted signal peptide region (i.e. from the N-terminal to the position immediately before the maximum of the Y-score). The bin size of the distribution is 0.02.

89% for prokaryotes (not divided into Gram-positive and -negative). When the original weight matrix is applied to our recent data set, however, the performance is much lower. This suggests a larger variation in the examples of the signal peptides found since then. It may, of course, also reflect a higher occurrence of errors in our automatically selected data than in the manually selected 1986 set.

In order to compare the strength of the neural network approach to the weight matrix method, we recalculated new weight matrices from our new data and tested the performances of these (results not shown). The weight matrix method was comparable to the neural networks when calculating the C-score, but was practically unable to solve the S-score problem

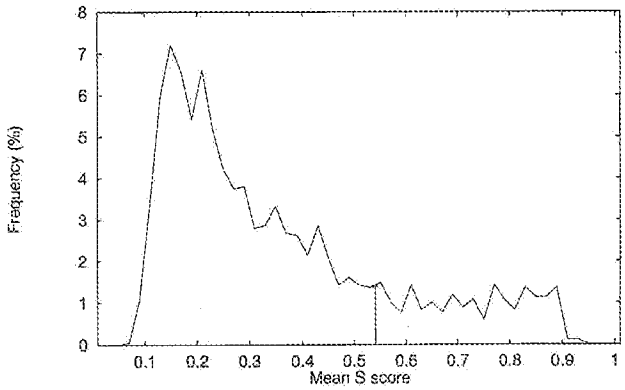


Fig. 4. Distribution of the mean signal peptide score (S-score) for all the predicted *H.influenzae* coding sequences. The mean S-score is calculated using networks trained on the Gram-negative data set. The bin size of the distribution is 0.02. The arrow shows the optimal cut-off for predicting a cleavable signal peptide. The predicted number of secretory proteins in *H.influenzae* (corresponding to the area under the curve to the right of the arrow) is 330 out of 1680 (20%).

and therefore did not provide the possibility of calculating the combined Y-score.

Note that the prediction performances reported here correspond to minimal values. The test sets in the cross-validation have a very low sequence similarity; in fact, the sequence similarity is so low that the correct cleavage sites cannot be found by alignment (Nielsen *et al.*, 1996a). This means that the prediction accuracy on sequences with some similarity to the sequences in the data sets will in general be higher.

The differences between the signal peptides from different organisms are apparent from Figure 1. The signal peptides from Gram-positive bacteria are considerably longer than those of other organisms, with much more extended h-regions, as observed previously (von Heijne and Abrahmsén, 1989). The prokaryotic h-regions are dominated by Leu (L) and Ala (A) in approximately equal proportions and in the eukaryotes they are dominated by Leu with some occurrence of Val (V), Ala, Phe (F) and Ile (I). Close to the cleavage site, the (–3,–1) rule is clearly visible for all three data sets, but while a number of different amino acids are accepted in the eukaryotes, the prokaryotes accept alanine almost exclusively in these two positions. In the first few positions of the mature protein (downstream of the cleavage site) the prokaryotes show certain preferences for Ala, negatively charged (D or E) amino acids, and hydroxy amino acids (S or T), while no pattern can be seen for the eukaryotes. In the leftmost part of the alignment, the positively charged residue Lys (K) [and to a smaller extent Arg (R)] is seen in the prokaryotes, while the eukaryotes show a somewhat weaker occurrence of Arg (barely visible in the figure) and almost no Lys. This corresponds well with the hypothesis that positive residues are required in the n-region where the N-terminal Met is formulated for prokaryotes, but not necessarily for eukaryotes where the N-terminal Met in itself carries a positive charge (von Heijne, 1985).

The difference in structure is reflected in the performances of the trained neural networks (see Table I). Gram-negative cleavage sites have the strongest pattern—i.e. the highest information content—and, consequently, they are the easiest to predict, both at the single-position and at the sequence level. The eukaryotic cleavage sites are significantly more difficult

to predict. Gram-positive cleavage sites are slightly more difficult to predict than the eukaryotic ones, which would not be expected from the sequence logos (Figure 1), since they show nearly as high an information content as the Gram-negative cleavage sites, but the longer Gram-positive signal peptides means that the cleavage sites have to be located against a larger background of non-cleavage site positions. The discrimination of signal peptides versus non-secretory proteins, on the other hand, is better for the eukaryotes than for the prokaryotes. This may be due to the more characteristic leucine-rich h-regions of the eukaryotic signal peptides.

The logos for the human and *E.coli* data sets are not shown, since they show no significant differences from those of the eukaryotes or Gram-negative bacteria respectively. Accordingly, the predictive performance was not improved by training the networks on single-species data sets. On the contrary, the *E.coli* signal peptides are predicted even better by the Gram-negative networks than by the *E.coli* networks (probably due to the relatively small size of the *E.coli* data set). In other words, we have found no evidence for species-specific features of the signal peptides of humans and *E.coli*.

Signal anchors often have sites similar to signal peptide cleavage sites after their hydrophobic (transmembrane) region. Therefore, a prediction method can easily be expected to mistake signal anchors for peptides. In Figure 3, the distribution of the mean S-score for the 97 eukaryotic signal anchors is included. It shows some overlap with the signal peptide distribution. If the standard cut-off of 0.5 is applied to the signal anchor data sets, 50% of the eukaryotic signal anchor sequences are falsely predicted as signal peptides (the corresponding figure for the human signal anchors is 75% when using human networks and 68% when using eukaryotic networks). With a cut-off optimized for signal anchor versus signal peptide discrimination (0.62), we were able to lower this error rate to 45% for the eukaryotic data set. The mean S-score still gives a better separation than the maximal C- or Y-score, which indicates that the pseudo-cleavage sites are in fact rather strong.

However, the pseudo-cleavage sites often occur further from the N-terminal than genuine cleavage sites do. If we do not accept signal peptides longer than 35 residues (this will exclude only 2.2% of the eukaryotic signal peptides in our data set), the percentage of false positives among the signal anchors drops to 28% for the eukaryotic and 32% for the human signal anchors (39% when using eukaryotic networks). When taking this into account, our method does provide a reasonably good discrimination between signal peptides and signal anchors. This has not been reported by any of the earlier published methods for signal peptide recognition.

Scanning the Haemophilus influenzae genome

We have applied the prediction method with networks trained on the Gram-negative data set to all the amino acid sequences of the predicted coding regions in the *Haemophilus influenzae* genome. The distribution of the mean S-score (from position 1 to the position with a maximal Y-score) is shown in Figure 4.

When applying the optimal cut-off value found for the Gram-negative data set, we obtained a crude estimate of the number of sequences with cleavable signal peptides in *H.influenzae*: 330 out of 1680 sequences or approximately 20%. If the maximal S-score is used instead of the mean S-score, the estimate comes out as 28% and with the maximal Y-score it is 14% (distributions not shown). If all three criteria are applied together, leaving only 'typical' signal peptides, we obtain 188 sequences (11%).

Some of the sequences predicted to be signal peptides according to the S-score but not according to the Y-score may be signal anchor-like sequences of type II (single-spanning) or type IV (multispanning) membrane proteins. This hypothesis is strengthened by a hydrophobicity analysis of the ambiguous examples (results not shown). If we apply the slightly higher cut-off optimized for the discrimination of signal anchors versus signal peptides in eukaryotes (0.62) to the mean S-score, the estimate is lowered from 20 to 15%.

On the other hand, some of the sequences predicted to be signal peptides according to the maximal Y-score but not the mean S-score may be the effect of the initiation codon of the predicted coding region having been placed too far upstream. In this case, the apparent signal peptide becomes too long and the region between the false and the true initiation codon will probably not have signal peptide character, thereby bringing the mean S-score of the erroneously extended signal peptide region below the cut-off. This is strengthened by the finding that these ambiguous examples are longer than average and contain more methionines.

In conclusion, we estimate that 15–20% of the *H.influenzae* proteins are secretory. However, a whole-genome analysis like this would be more reliable if combined with other analyses, notably transmembrane segment predictions and initiation site predictions.

Method and data publicly available

The finished prediction method is available both via an e-mail server and a WWW server. Users may submit their own amino acid sequences in order to predict whether the sequence is a signal peptide and, if so, where it will be cleaved. We recommend that only the N-terminal part (say 50–70 amino acids) of the sequences is submitted, so that the interpretation of the output is not obscured by false positives further downstream in the protein.

The user is asked to choose between the network ensembles trained on data from Gram-positive, Gram-negative or eukaryotic organisms. We did not include the networks trained on the single-species data sets in the servers, since these did not improve the performance.

The values of the C-, S- and Y-scores are returned for every position in the submitted sequence. In addition, the maximal Y-score, maximal S-score and mean S-score values are given for the entire sequence and compared with the appropriate cut-offs. If the sequence is predicted to be a signal peptide, the position with the maximal Y-score is mentioned as the most likely cleavage site. A graphical plot in postscript format, similar to those in Figure 2, may be requested from the servers. We strongly recommend that a graphical plot is always used for the interpretation of the output. The plot may give hints about, for example, multiple cleavage sites or erroneously assigned initiation, which would not be found when using only the maximal or mean score values.

The address of the mail server is signalp@cbs.dtu.dk. For detailed instructions, send a mail containing the word 'help' only. The WWW server is accessible via the Center for Biological Sequence Analysis homepage at http://www.cbs.dtu.dk/.

All the data sets mentioned in Table I are available from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. Retrieve the file README for detailed descriptions of the data and the format.

The FTP server and the mail server can both be accessed directly from the WWW server.

References

Arrigo,P., Giuliano,F., Scalia,F., Rapallo,A. and Damiani,G. (1991) *CABIOS*, **7**, 353–357.

Bairoch,A. and Boeckmann,B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol.*, **220**, 49–65.

Fleischmann,R. *et al.* (1995) *Science*, **269**, 449–604.

Gierasch,L.M. (1989) *Biochemistry*, **28**, 923–930.

Hirst,J.D. and Sternberg,M.J.E. (1992) *Biochemistry*, **31**, 7211–7218.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.

Ladunga,I., Czakó,F., Csabai,I. and Geszti,T. (1991) *CABIOS*, **7**, 485–487.

Mathews,B. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.

Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996a) *Proteins*, **24**, 165–177.

Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1997) *Int. J. Neural Sys.*, in press.

Presnell,S.R. and Cohen,F.E. (1993) *Annu. Rev. Biophys. Biomol. Struct.*, **22**, 283–298.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Rapoport,T.A. (1992) *Science*, **258**, 931–936.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) In Rumelhart,D., McClelland,J. and the PDP Research Groups (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA, pp. 318–362.

Schneider,G. and Wrede,P. (1993) *J. Mol. Evol.*, **36**, 586–595.

Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.

Shannon,C.E. (1948) *Bell System Technol. J.*, **27**, 379–423, 623–656.

von Heijne,G. (1983) *Eur. J. Biochem.*, **133**, 17–21.

von Heijne,G. (1985) *J. Mol. Biol.*, **184**, 99–105.

von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

von Heijne,G. (1988) *Biochim. Biophys. Acta*, **947**, 307–333.

von Heijne,G. (1990) *J. Membrane Biol.*, **115**, 195–201.

von Heijne,G. and Abrahmsén,L. (1989) *FEBS Lett.*, **244**, 439–446.

Received April 19, 1996; revised September 2, 1996; accepted September 12, 1996

Identifying the CDRs

This protocol describes how to identify the CDRs (Kabat definition) by examining the sequence. Of course there are always (minor) exceptions to these rules, so the word 'always' should be interpreted with care! For example, CDR-L2 is always 7 residues, but antibody NEW (Protein Databank code: 7FAB, http://www.rcsb.org) has a deletion in this region. This also means that the position of the start of CDR-L3 is no longer 33 residues after the end of CDR-L2.

CDR-L1

- **Start** Approx residue 24
- **Residue before** is always C
- **Residue after** is always W. Typically WYQ, but also, WLQ, WFQ, WYL
- **Length** 10 to 17 residues

CDR-L2

- **Start** Always 16 residues after the end of CDR-L1
- **Residues before** generally IY, but also, VY, IK, IF
- **Length** always 7 residues

CDR-L3

- **Start** Always 33 residues after end of CDR-L2
- **Residue before** is always C
- **Residues after** always FGXG
- **Length** 7 to 11 residues

CDR-H1

- **Start** Approximately residue 31 (always 9 after a C) (Chothia/AbM definition starts 5 residues earlier)
- **Residues before** always CXXXXXXXX
- **Residues after** always W. Typically WV, but also WI, WA
- **Length** 5 to 7 residues (Kabat definition); 7 to 9 residues (Chothia definition); 10 to 12 residues (AbM definition)

CDR-H2

- **Start** Always 15 residues after the end of Kabat/AbM definition of CDR-H1
- **Residues before** typically LEWIG, but a number of variations
- **Residues after** K[RL]IVFT[AT]SIA (where residues in square brackets are alternatives at that position)
- **Length** Kabat definition 16 to 19 residues (AbM definition and most recent Chothia definition ends 7 residues earlier; earlier Chothia definition starts 2 residues later and ends 9 earlier)

CDR-H3

- **Start** Always 33 residues after end of CDR-H2 (always 3 after a C)
- **Residues before** always CXX (typically CAR)
- **Residues after** always WGXG
- **Length** 4 to 24 residues

Screening new antibody sequences

Given a new antibody sequence, one is likely to wish to assign families and subgroups using the tools described above. An additional facility is available at http://www.bioinf.org.uk/abs/seqtest.html to identify unusual features in the sequence.

It is simply necessary to enter the amino acid sequence of your Fv fragment (one or both chains). Optionally you may include the whole Fab fragment, but only the Fv portion will be tested.

The tool aligns the provided sequence with a standard sequence in order to assign standard Kabat numbering and then uses the KabatMan database to identify unusual amino acids (i.e. those occurring in less than 1% of the data in the database). This allows the identification of potential cloning artifacts and sequencing errors. If unusual features are verified as being correct, then these residues are likely to be critical to the specificity of the antibody. The method is described in detail at http://www.bioinf.org.uk/abs/seqmethod.html.

The results need to be examined carefully. A typical sequence has 1-2 'unusual' residues. Very unusual sequence features and loops longer than anything observed in the current Kabat database may cause the alignment